Statistical interference in high dimensions and applications to medical data

Sheikh, Mansoor

Awarding institution:
King's College London

Statistical Inference in High Dimensions and Applications to Medical Data

Mansoor Sheikh

Department of Mathematics
King's College London

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2019

Acknowledgements

I am grateful to my supervisor, Professor Ton Coolen, for his encouragement over the last four years, for the excellent research environment he provided and for unlimited amounts of good coffee. His experience has gotten me out of trouble more than a few times. Most of all, Ton has taught the value of patience.

This thesis would be half what it is without the generous help of everyone in the Institute for Mathematical and Molecular Biomedicine. A special thanks to Alexander Mozeika and Fabián Aguirre López for always being there to discuss the minutiae of statistical inference, physics and politics. My thanks are also due to Peter Young and Mark Rowley for reviewing sections of my thesis and to my upgrade examiners, Professor Reimer Kühn and Dr Pierpaolo Vivo, for constructive comments. Finally, I would like to thank my family (particularly Lara) for putting up with me and my parents for instilling in me the virtues of education.

I would also like to extend my gratitude to the Biotechnology and Biological Sciences Research Council (award 1668568) and GSK Ltd for their generous financial support throughout my PhD.

Publications

Sections of this thesis have been based on the following publications.

1. Sheikh, M. and Coolen, ACC. (2019). Analysis of overfitting in the regularized Cox model. J Phys A-Math Theor, 52(38):384002

2. Sheikh, M. and Coolen, ACC. (2020). Accurate Bayesian data classification without hyperparameter cross-validation. J Classif, 37(2):277-297

3. Coolen, ACC., Sheikh, M. et al. (2020). Replica analysis of overfitting in generalized linear models. J Phys A-Math Theor, 53(36):365001

4. Santaolalla, A., Sheikh, M. et al (2018). Improved resection margins in breast-conserving surgery using Terahertz Pulsed imaging data (under review).

   https://arxiv.org/abs/1805.01349

Declaration

I declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

This dissertation is my own work. Chapters 4 and 6 are the result of joint publications in peer-reviewed journals with my supervisor, Professor ACC Coolen. Chapter 7 is the outcome of work done in collaboration with others (see Publications section) and is currently under peer-review.

This dissertation contains fewer than 100,000 words including appendices, bibliography, footnotes, tables and equations.

Mansoor Sheikh
December 2019

Abstract

The message from medicine is clear. We are in possession of vast amounts of data from sources such as electron microscopes, magnetic resonance imaging and DNA microarray technology. The aim is to translate this into a world of faster drug discovery, more accurate automated diagnosis and early warnings of impending disease. Statistics, and in particular statistical inference, has a key role to play in providing a principled approach to this task. In practice, solid theoretical underpinnings are often replaced with heuristics suited only to the data-set in hand. Traditional statistical tools are often ill-equipped to cope with the structure of this new flood of data. The need for methodological improvements presents a major opportunity to progress medical sciences.

The scientific method relies on carefully designed experiments, typically collecting N samples each with a small number p of measurements, in order to test a hypothesis or to discover a causal relationship. When p is much smaller than N, existing statistical methods are satisfactory for this purpose. Today, however, the approach is quite different. Extensive quantities of data are available due to the decreasing cost of high throughput measurement devices and data storage alongside rapidly increasing computational power. Examples from biomedicine include genetic and epigenetic data as well as the growing availability of real-time health information from wearable devices and hospital intensive care units. In these cases, the data dimension p can be comparable to the number of samples N. The task is now to look for patterns or correlations in this data without first providing a plausible hypothesis. The response to this problem has been to search for a lower dimensional representation of the data. To achieve this, methods of variable selection, regularization or projection are routinely used.

Given its wide-reaching implications for biomedical data studies we examine the phenomenon of overfitting, one of the primary challenges of high-dimensional inference, from two different perspectives: statistical physics in Part I and Bayesian inference in Part II. In the first approach, we consider all variables of the dataset and investigate the inference outcomes in the regime where p is comparable to N. Surprisingly, for a family of models, we have found systematic and reproducible effects which are a function of the ratio p/N.

These always act to reduce prediction accuracy when applied to unseen data samples but given their predictable behaviour, can be corrected for. We find a relationship between the following: bias of both the mean and variance of inferred regression outcomes; the ratio $\zeta = p/N$ and the amount of regularization η. In addition, hypothesis tests and confidence interval estimation, being usually based on asymptotic results derived for fixed p, become increasingly inaccurate with increasing ζ. These effects are investigated for linear, logistic and Cox regression by using an information-theoretic overfitting measure and the replica method of statistical physics. In the process, we gain a better understanding of the inference under ML, MAP and class-imbalanced data. In Part II, we use the more established route of Bayesian analysis for the purpose of statistical inference. We find, that with judicious model setup, the family of integrable Bayesian models is extended resulting in a novel model branch with superior performance under certain conditions. This is pursued in an attempt to reduce overfitting when applied to classification problems. This classifier is successfully applied to a medical data study for detecting breast cancer in ex vivo patient samples.

In sum, the regime of $p \sim N$ is not only of academic concern. The collection of large datasets generated by modern biomedical applications, particularly in the post-genomic era, are a necessary starting point; the statistical tools developed in this thesis may elaborate possible ways to draw or sharpen conclusions from this data.

Notation

Some notation and symbols are collected here for convenience.

- $|\mathbf{M}|$ or Det $\mathbf{M}$ represents the determinant of matrix $\mathbf{M}$
- $\mathrm{Tr}\,\mathbf{M} = \sum_\mu M_{\mu\mu}$ represents the trace of matrix $\mathbf{M}$
- $\mathbf{M} \succ 0$ means matrix $\mathbf{M}$ is a positive definite matrix.
- The modulus of a p-dimensional vector β is denoted as $|\beta| = \sqrt{\beta_1^2 + \cdots + \beta_p^2}$
- $\langle f(x) \rangle$ represents the average of $f(x)$ with respect to a probability distribution p(x). It is equivalent to the expectation $\mathbb{E}[f(x)] = \int \mathrm{dx}\ \mathrm{p(x)}\ \mathrm{f(x)}$ for continuous random variables or $\mathbb{E}[f(x)] = \sum_{i=1}^{N} p(x_i)\ f(x_i)$ for the discrete case.
- The short-hand for integration over a Gaussian measure $\mathrm{D}z = (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}z^2} \mathrm{d}z$ over the real line.
- $\mathrm{N}(\mu, \sigma^2)$ or $\mathrm{N}(t|\mu, \sigma^2)$ represents a normal or Gaussian distribution with mean μ and variance σ^2. We denote the p-dimensional multivariate equivalent as $\mathrm{N}_p(\mu, \Sigma)$
- $\mathrm{U}(0,1)$ represents a uniform distribution with domain $[0,1]$
- Bracketed numbers generally refer to equations e.g. (2.29)
- $\delta(\ldots)$ represents the Dirac delta function and $\delta_{(\ldots)}$ the Kronecker delta function.
- $\mathbb{1}$ is the identity matrix i.e. $I_{\mu\nu} = \delta_{\mu\nu}$
- log represents the natural logarithm.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

We will first review the foundations of statistical inference with data consisting of N samples, each with p associated measurements (or dimensions), such as N medical patients with their associated gene expression data ($p \sim 10^4 - 10^5$) [21, 24, 73]. Time-series data is not covered in this thesis. By considering the regime where $N < p$, we find that some of the assumptions underlying traditional methods no longer hold leading to biased or inaccurate results [8, 16, 44]. We address the seemingly counter-intuitive problem of having too many measurements per patient.

1.1 What is statistical inference?

The process of statistical inference begins with the estimation of model parameters from a data sample. This generally takes the form of point estimates of relevant parameters along with a variance measure. Conclusions can be drawn using either confidence intervals or hypothesis tests, both of which require sampling distributions of their respective test statistics [67]. For example, the log-likelihood statistic (or deviance), commonly used in hypothesis testing, is known to have a χ^2 distribution [37].

Inference in parametrized models typically assumes knowledge of the data-generating process [49]. For example, we could assume a linear relationship such as $t = \beta \cdot \mathbf{z} + \varepsilon, \;\; \varepsilon \sim \mathrm{N}(0, \sigma^2)$ where $t \in \mathbb{R}$ is the response variable, $\mathbf{z}_i \in \mathbb{R}^p$ are the covariates and $\varepsilon_i \sim \mathrm{N}(0, (\sigma^0)^2)$ represents Gaussian noise with variance $(\sigma^0)^2 > 0$. The model choice often relies on convention or on the grounds of analytical tractability rather than closeness to scientific truth. In fact, there are rarely laws, akin to the laws of physics, applicable to biomedical data. This use of empirical models along with noisy measurements and heterogeneous biological sources means it is unlikely that our assumed model is the true one. In this thesis, the issue of model misspecification is only touched upon in the interests of developing a plausible theory.

Having made a model choice, we can write a joint distribution of the data conditioned on the parameter set ϑ. Further assuming each sample is independently sampled from the population distribution, we can factorize this joint distribution.

$$p(t_1,\dots,t_N;z_1,\dots,z_N|\vartheta) = \prod_{i=1}^{N} p(t_i;\mathbf{z}_i|\vartheta) \tag{1.1}$$

Here the dataset $\mathscr{D} = \{(t_i,\mathbf{z}_i)\}_{i=1}^N$ contains the random variables and the parameter vector ϑ is fixed. The likelihood function $L(\vartheta)$, which is algebraically equivalent to (1.1), considers ϑ as the random variable and fixed $\mathscr{D}$. It quantifies the probability of $\mathscr{D}$ occurring given a set of model parameter values. To estimate values for ϑ, the Maximum Likelihood Estimator (MLE) is commonly used. This is the parameter value which maximizes the likelihood function over the parameter space Θ

$$\hat{\vartheta}_{\mathrm{ML}} = \underset{\vartheta\in\Theta}{\operatorname{argmax}}\ L(\vartheta) \tag{1.2}$$

where the hat notation e.g. $\hat{\vartheta}$ denotes an inferred parameter. Equivalently, we can maximize the log likelihood function $\ell(\vartheta) = \log L(\vartheta)$ since log is a monotonically increasing function of its argument. This approach has dominated inference methods since its introduction by R.A. Fisher [49]. Other methods using classical (or frequentist) tools [74] include Least Squares estimation and the Method of Moments. In some cases, the MLE can be found analytically by setting the derivative of the log likelihood to zero

$$U(\hat{\vartheta}_{\mathrm{ML}}) = \frac{\partial \ell(\vartheta)}{\partial \vartheta} = 0 \tag{1.3}$$

where $U(\vartheta)$ is the so-called score function. For a correctly specified linear regression model, the residuals are normally distributed and the MLE coincides with the Least Squares estimator $\hat{\vartheta}_{\mathrm{LS}} = \sum_{i=1}^{N}(t_i - \beta\cdot\mathbf{z}_i)^2$. When no analytical solution to (1.3) is possible, numerical methods are required. These often depend on the eigenvalues of the FIM, see (1.4) and [5], with the minimization slowing down around small values corresponding to flat regions of the likelihood function.

To make the idea concrete, we plot the log likelihood for a discrete random variable with probability mass function $f(k) = \frac{N!}{(N-k)!\,k!}\beta^k(1-\beta)^{N-k}$ where N is the number of trials, $k\in\mathbb{Z}^+$ is the number of successes and the scalar parameter β represents the probability of success for each trial. As N increases for this binomial model, we see from Figure 1.1 that the maximum value of the likelihood function tends towards the true parameter value $\beta^0 = 0.5$. The uncertainty around this maximum value is given by the curvature $\frac{\partial^2\ell}{\partial\beta^2}$ or $\frac{\partial U}{\partial\beta}$, which for

multivariate models, is defined by the eigenvalues of the Fisher Information Matrix (FIM).

$$[\mathbf{I}(\beta)]_{\mu\nu} = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \beta_\mu \beta_\nu}\right] \tag{1.4}$$

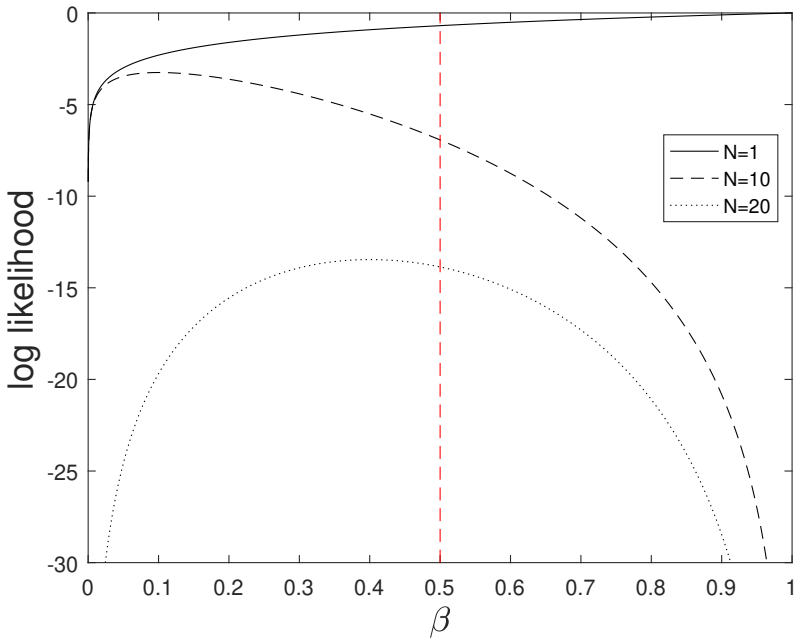

Fig. 1.1 The log likelihood function for a binomial model with true parameter $\beta^0 = 0.5$ and $N = \{1, 10, 20\}$.

Next we describe an equivalent procedure to maximizing the likelihood which will be useful in the following chapters. The Kullback-Leibler (KL) divergence between probability distributions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \int \mathrm{dx}\ \mathrm{p(x)}\ \log\ \frac{\mathrm{p(x)}}{\mathrm{q(x)}} \tag{1.5}$$

where support$(p) \subseteq$ support(q). It is not a metric since the triangle inequality does not hold and it is not symmetric in its arguments i.e. $D(p\|q) \neq D(q\|p)$. The lower bound $D(p\|q) \geq 0$ is achieved at $p(x) = q(x)$. Consider the KL divergence between an empirical data distribution $\hat{P}_{\mathscr{D}}(x) = N^{-1}\sum_{i=1}^{N} \delta(x - x_i)$ and a parametrized distribution $P_\vartheta(x)$

$$\begin{aligned} D(\hat{P}_{\mathscr{D}}\|P_\vartheta) &\equiv \int \mathrm{dx}\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x})\ \log\ \frac{\hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x})}{\mathrm{P}_\vartheta(\mathrm{x})} \\ &= \int \mathrm{dx}\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x})\ \log\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x}) - \int \mathrm{dx}\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x})\ \log\ \mathrm{P}_\vartheta(\mathrm{x}) \end{aligned} \tag{1.6}$$

The parameterized distributions chosen in this thesis have a non-zero probability measure everywhere on $\mathbb{R}^p$ and so fulfill the support condition. Minimizing the KL divergence is equivalent to maximizing $\int \mathrm{dx}\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x}) \log \mathrm{P}_{\vartheta}(\mathrm{x})$ with respect to ϑ

$$\begin{aligned}\int \mathrm{dx}\ \hat{\mathrm{P}}_{\mathscr{D}}(\mathrm{x}) \log \mathrm{P}_{\vartheta}(\mathrm{x}) &= \int \mathrm{dx}\ \left[\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta(\mathrm{x}-\mathrm{x}_{\mathrm{i}})\right] \log \mathrm{P}_{\vartheta}(\mathrm{x}) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left[\int \mathrm{dx}\ \delta(\mathrm{x}-\mathrm{x}_{\mathrm{i}}) \log \mathrm{P}_{\vartheta}(\mathrm{x})\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \log P_{\vartheta}(x_i) = \frac{1}{N} \log \left[\prod_{i=1}^{N} P_{\vartheta}(x_i)\right]\end{aligned} \tag{1.7}$$

The last term is the log likelihood and hence minimization of the KL divergence is equivalent to maximum likelihood estimation.

Having reviewed the frequentist approach to statistical inference, we briefly consider Bayesian methods which form the subject of Part II of this thesis. So far the key object in inference is the likelihood function $p(\mathscr{D}|\vartheta)$ and maximization of this function results in the *true* parameter values (at least asymptotically). In the Bayesian approach, it is acknowledged that there are a range of parameter values consistent with the data [75]. This uncertainty is incorporated through a prior distribution and Bayes Theorem enables a posterior probability distribution $p(\vartheta|\mathscr{D})$ to be calculated

$$p(\vartheta|\mathscr{D}) = \frac{p(\mathscr{D}|\vartheta)\ p(\vartheta)}{p(\mathscr{D})} \tag{1.8}$$

Apart from the likelihood function $p(\mathscr{D}|\vartheta)$, the other terms did not appear in the frequentist version of inference so we will briefly identify them here. Prior information is incorporated into the model via $p(\vartheta)$. Prior distributions can be categorized into conjugate [58, 82], Jeffrey's [76], maximum entropy [74], reference [12, 146] and non-conjugate priors amongst others. Setting $p(\vartheta)$ to a uniform distribution recovers the maximum likelihood case. We have omitted conditioning on the parameters of the prior distribution, the so-called hyper-parameters $\mathscr{H}$ [91], for clarity. Their role will be explained in detail in Part II. Finally the evidence term in the denominator normalizes the posterior probability distribution and is defined as

$$p(\mathscr{D}) = \int \mathrm{d}\vartheta\ p(\mathscr{D}|\vartheta)\ p(\vartheta) \tag{1.9}$$

Inference using the maximum value of this posterior distribution is termed *maximum a posteriori* (MAP) inference. Integrating over ϑ results in a fully Bayesian approach to

inference. Part II defines these terms in more detail and applies Bayesian inference to a classification problem.

To progress with the inference procedure, we are required to make a number of assumptions. When these are invalid the accuracy of our results is reduced. Asuming perfect knowledge of the true data-generating process is clearly an idealized scenario especially in the biomedical setting and cases. Where it is not true, we have model mismatch or misspecification. Further, even if the distribution was known, it may change over time. We now introduce problems specific to the regime where the number of dimensions p is comparable to the sample size N.

1.1.1 Overfitting

Overfitting is a major obstacle to accurate inference and occurs when there are insufficient data samples to characterize each dimension (see e.g. [4, 26, 61, 133] for examples from logistic regression, gamma distributions and Cox models). It does not seem to have a clear definition but is typically thought of as fitting model parameters too closely to the empirical data. This data contains the true signal along with measurement and sampling noise.

We define the success of the inference procedure by an error function $\varepsilon(\vartheta;\mathbf{z}_i,t_i)$ which is non-negative and is equal to zero if the response variable matches the true result. For real-valued response variables, this is often chosen to be a quadratic function. The training error is the sum of this function over the N samples of the training data $\mathscr{D}$. Following [121],

$$E_T(\vartheta) = \frac{1}{N}\sum_{i=1}^{N} \varepsilon(\vartheta;\mathbf{z}_i,t_i) \tag{1.10}$$

However the aim of inference is often to predict the response variable for unseen data samples. Therefore the generalization error is calculated by averaging over the entire data distribution

$$E_G(\vartheta) = \int \mathrm{d}\mathbf{z}\ \mathrm{dt}\ \mathrm{p}(\mathbf{z},\mathrm{t})\ \varepsilon(\vartheta;\mathbf{z},\mathrm{t}) \tag{1.11}$$

A fingerprint of overfitting is the increasing difference between $E_T(\vartheta)$ and $E_G(\vartheta)$. For example, see Figures 6.5-6.6. Before addressing approaches to reducing overfitting, we give some background to the problem of inference in high-dimensions.

1.2 The problem with high-dimensional data

Let us start by defining what we mean by high-dimensional data. Each sample of a dataset can be represented as a data point $\mathbf{z}_i \in \mathbb{R}^p$ with components $(z_1, z_2, \ldots, z_p)$ in a p-dimensional vector space. As p increases, it is not straightforward to represent the data graphically and much of the intuition gained from two- or three- dimensional space no longer holds. The appropriate inference methodology often depends on the structure of the data and it is useful to define three broad categories following [140].

- *Classical asymptotics*. Here the model (and hence p) is fixed and the number of samples, $N \to \infty$. Statistical estimators typically become asymptotically unbiased under these conditions (see for example [95]).

- *High-dimensional asymptotics*. This regime, also known as the *Kolmogorov regime*, is defined by $p, N \to \infty$ and $p/N \sim \mathcal{O}(1)$. The ratio $\zeta \equiv p/N \in (0, \infty)$ will be used to characterize the inference problem. Early examples of high-dimensional asymptotic results can be found in [109]. More recently statistical physics methods have been applied to inference problems [97, 86]. These rely on taking the thermodynamic limit $N \to \infty$ and naturally fall into the regime of high-dimensional asymptotics.

- *High-dimensional bounds*. As the name suggests, these methods derive mathematical bounds on statistics as a function of finite p and N. Recent introductions to the subject can be found in [17, 138, 140].

We now give an overview of classical and high-dimensional asymptotics and the related fields of feature selection and regularization since these are most relevant to this thesis.

1.2.1 Classical asymptotics

The MLE has a number of useful properties in the limit of large N (and fixed p) such as consistency: $\hat{\vartheta}_{\mathrm{ML}} \xrightarrow{p} \vartheta^0$ and asymptotic normality: $\sqrt{N}(\hat{\vartheta}_{\mathrm{ML}} - \vartheta^0) \rightsquigarrow \mathrm{N}(0, \mathbf{I}^{-1}(\vartheta))$ where ϑ^0 is the true parameter vector, $\xrightarrow{p}$ means convergence in probability and $\rightsquigarrow$ means convergence in distribution. However within a few decades of its introduction, there was a realization that this estimator is biased when sample size is small (see [6, 7] and more recently [34, 48]). The source of this bias originates from the curvature of the score function and from the accumulation of errors from each of the p components which becomes material when $p \sim \mathcal{O}(N)$. Attempts to correct for this bias were either analytical or numerical. Analytical

methods can be further characterized as either "corrective" or "preventative" in the language of [48].

Corrective analytic methods. The MLE is calculated along with a correction factor. In 1953, Bartlett gave first order bias corrections for one parameter [6] and many parameters [7]. This was followed by a number of papers using higher order cumulants of the MLE (for example [67]). These ideas were developed in a series of papers [4, 35, 125, 126]. Note that corrective methods become problematic if the MLE does not exist.

Preventative analytic methods. The modified score function approach of [48] is used to correct for the $\mathscr{O}(N^{-1})$ bias term. This is done via an expansion of the score function and using the techniques of cumulants to find the bias term.

Numerical methods. These methods re-sample the original dataset to produce multiple samples which are used to calculate the bias and the standard error. The Jackknife method [108, 137] uses sampling without replacement and the bootstrap [39, 40] uses sampling with replacement.

1.2.2 High-dimensional asymptotics

This regime is defined by $p,N \to \infty$ and $p/N \sim \mathscr{O}(1)$. We have hinted at data with $p \sim N$ causes problems in classical inference methods but what relevance does the limit $p,N \to \infty$ have? To motivate this, interesting behaviour of the regression outcomes is illustrated through simulations results in Figures 2.1, 3.1, 4.2 for the linear, logistic and Cox regression models. See [37, 95] for an introduction to these models. As well as asymptotic limits tending to simplify calculations, we find, through simulations, that our results are valid for relatively small values of $p \sim 10^3$ which is within the regime of modern medical data.

The difficulties presented in this regime can be illustrated by the important problem of estimating population eigenvalues from empirical data. We choose this example since empirical covariance matrices frequently appear in multivariate statistical analysis. For example, consider the use of Principal Component Analysis (PCA), an important tool in the analysis of high-dimensional data [70, 79]. Rather than estimating all $p(p+1)/2$ entries of a covariance matrix, PCA estimates the p eigenvalues in order to reduce the dimension of the data [79].

Consider data $\mathbf{z}_i^T = (z_{i1},\ldots,z_{ip})$ where $i = \{1,2,\ldots,N\}$ forming a data matrix generated from a model with population mean $\mu = (\mu_1,\ldots,\mu_p)^T \in \mathbb{R}^p$ and population covariance

matrix $\Sigma \in \mathbb{R}^{p\times p}$.

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_N^T \end{pmatrix} = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ldots & \vdots \\ z_{N1} & \cdots & z_{Np} \end{pmatrix} = \mathbb{R}^{N\times p} \tag{1.12}$$

and the sample mean as

$$\langle \mathbf{z} \rangle = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} z_{i1} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} z_{ip} \end{pmatrix} = \begin{pmatrix} \langle z_1 \rangle \\ \vdots \\ \langle z_p \rangle \end{pmatrix} \tag{1.13}$$

The components of the sample covariance matrix $\hat{\Sigma}$ can be written as

$$\begin{aligned} \hat{\Sigma}_{jk} &= \frac{1}{N-1}\sum_{i=1}^{N} \left[z_{ij} - \langle z_j \rangle\right]\left[z_{ik} - \langle z_k \rangle\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N} \left[z_{ij} - \mu_j\right]\left[z_{ik} - \mu_k\right] - \frac{N}{N-1}\left[\langle z_j \rangle - \mu_j\right]\left[\langle z_k \rangle - \mu_k\right] \end{aligned} \tag{1.14}$$

Assuming the samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are i.i.d, we find as $N \to \infty$, $\hat{\Sigma}_{jk} \to \Sigma_{jk}$ i.e. the sample covariance matrix is a consistent estimator of the population covariance matrix. However, the problems that occur non-asymptotically can be seen in Figure 1.2 where the eigenvalues of $\hat{\Sigma}$ are plotted for $p = 100$, $N = \{50, 75, 100, 1000\}$ and $\Sigma = \mathbb{I}$, the $p \times p$ dimensional identity matrix. The sample eigenvalue mean is unity but the largest sample eigenvalues are too large and the smallest sample eigenvalues are too small compared to the population eigenvalue spectrum. To understand this phenomenon analytically, we make use of a central result of random matrix theory, the Marčenko-Pastur equation [94]. Consider an $N \times p$ matrix, $\mathbf{Z}$, with zero mean, unit variance independent entries. The limiting eigenvalue distribution, $\rho(\lambda)$, of the $p \times p$ empirical covariance matrix, $N^{-1}\mathbf{Z}^T\mathbf{Z}$ is

$$\rho(\lambda) = \frac{\sqrt{(\lambda - \lambda_{\min})(\lambda_{\max} - \lambda)}}{2\pi\zeta\lambda} \tag{1.15}$$

in the limit $p, N \to \infty, \quad p/N \to \zeta \leq 1$ where $\lambda_{\min} = \left(1 - \sqrt{\zeta}\right)^2$ and $\lambda_{\max} = \left(1 + \sqrt{\zeta}\right)^2$. Again, the resulting eigenvalue spectrum, displayed in Figure 1.3 for three non-zero values of ζ, is quite different from $\rho(\lambda) = \delta(\lambda - 1)$, the spectrum of the population covariance matrix $\Sigma = \mathbb{I}$. We can see from Figure 1.3 that the average eigenvalue is one and can confirm

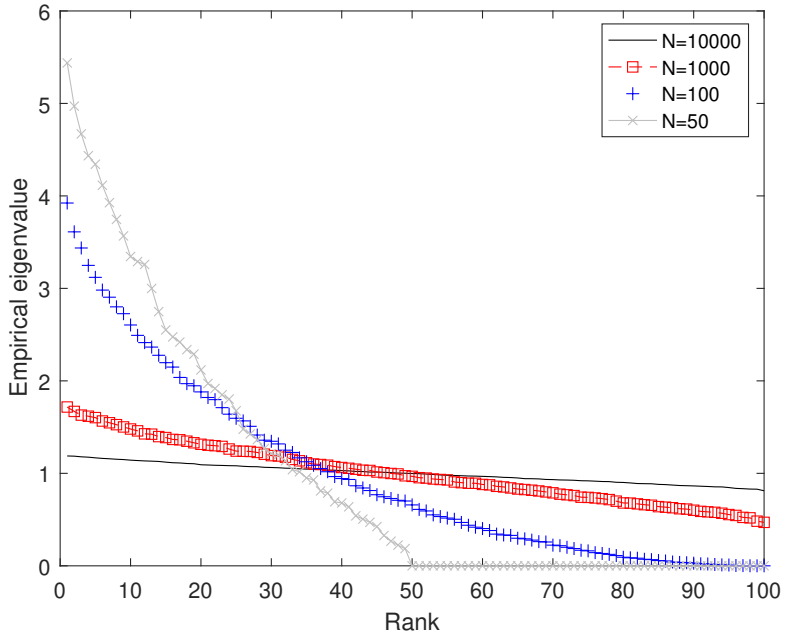

Fig. 1.2 Empirical eigenvalue spectra for various p/N ratios. For each value of $N = \{10000, 1000, 100, 50\}$, data was generated with distribution $\mathrm{N}_p(\mathbf{0}, \mathbb{I})$. The empirical eigenvalue spectra are plotted in descending order to illustrate the problems of eigenvalue estimation when $\zeta = p/N$ increases. As expected the spectra tends to $\delta(1)$ as $N \to \infty$ since the population covariance matrix is the identity.

this analytically by calculating the first moment of the Marčenko-Pastur distribution (see Appendix B.6.1).

This problem is not specific to the estimation of eigenvalues in high dimensions. We will find in the coming chapters that it is also present in the inference problem for generalized linear models where parameter estimates become increasingly biased as ζ increases. Mitigating this effect has focused on searching for low-dimensional representations of the data. One may believe that the true data-generating model is of a lower dimension than the experimental observations[1] or one may simply have a pragmatic desire to reduce the complexity of the data to allow for more efficient subsequent steps. This is achieved through regularization which either shrinks or eliminates regression coefficients [55, 56, 136] or by feature selection [2, 8, 43, 60] which finds an informative subset of variables.

[1] Gene expression data may contain $20,000$ measurements but only 100 of them have a recognizable biochemical pathway to the phenomenon of interest.

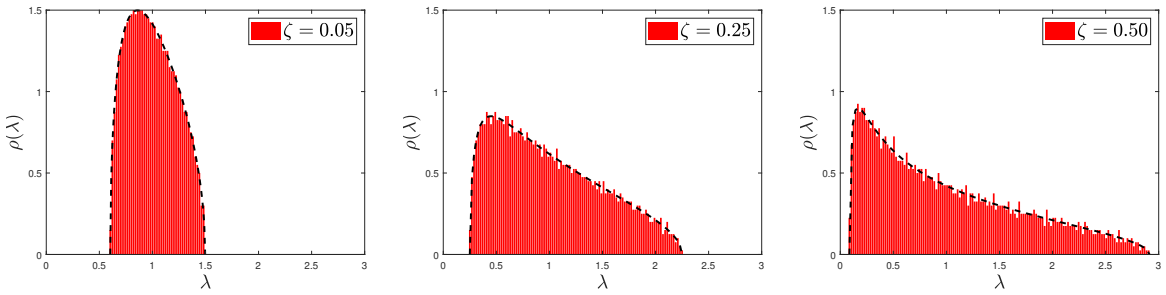

Fig. 1.3 Histogram of empirical eigenvalues for three values of ζ: 0.05 (left), 0.25 (centre) and 0.50 (right) with true covariance matrix $\Sigma = \mathbb{1}$. The solutions to the Marčenko-Pastur equations (1.15) are shown as black dashed lines.

1.2.3 Regularization methods

Diverging inference outcomes, a fingerprint of overfitting, led to the idea of adding a penalty term to the ML loss function suppressing the number or magnitude of the regression coefficients $\{\beta_\mu\}_{\mu=1}^p$ [56, 136]. Examples include penalties terms such as the sum of absolute parameter values (LASSO) $p(\beta) \propto \exp[-\eta \sum_{\mu=1}^p |\beta_\mu|]$ or the sum of squared parameter values (ridge regression) $p(\beta) \propto \exp[-\eta \sum_{\mu=1}^p \beta_\mu^2]$.

The penalty parameter η controls the amount of regularization and is often estimated by cross-validation [99]. This procedure randomly partitions the given dataset into training and validation sets. The model is trained on the former and its accuracy is measured on the latter. Since this is performed many times with different partitions, cross-validation is often time-consuming and computationally expensive for large datasets. In addition, it is wasteful of data since only a fraction is used for calibrating the model parameters. An alternative is leave-one-out cross-validation (LOOCV) [59, 132] where the model is trained on $N-1$ of the N samples with validation on the remaining data. This is repeated leaving out each data sample in turn.

In Part I, regularization is explicitly introduced in the model setup. In Part II, it is a natural outcome of introducing prior probabilities in Bayesian inference.

1.2.4 Feature selection

Feature selection (or variable selection) methods identify a subset of variables that are informative of outcomes [8, 43, 60]. This approach is intuitively appealing since some features may be irrelevant. The variables themselves are not altered unlike, for example, projection methods such as Principal Component Analysis [70]. By utilizing fewer variables in the inference method, feature selection typically requires less computational resource. The

resulting smaller subset of variables increases interpretability of results and can act as a guide to further experimental work by suggesting which features to investigate. Feature selection can be broadly categorized as

1. Filter methods. A statistical test is applied to variables either individually or in subsets without reference to the statistical inference method. This has the advantage of low computational expense.

2. Wrapper methods incorporate the inference method into the process. A subset of features is selected and subsequently the classification/regression method is assessed on this smaller number of features.

3. Embedded methods build feature selection into the inference method itself.

Bayesian variable selection [60] introduces a random vector $\gamma = \{0,1\}^p$ along with an associated prior probability distribution. Each component of γ indicates whether the corresponding variable should be included in the model. The prior can be conjugate or non-conjugate leading to Gibbs or Markov Chain Monte Carlo methods to sample from the posterior distribution.

1.3 Organization of thesis

This thesis is split into two parts, both of which examine the behaviour of supervised learning techniques when the data dimension p is of the same order as the sample size N.

Part I: Statistical Physics of High-Dimensional Inference. A general model examining inference outcomes in the regime $p/N > 0$ is developed using the framework of equilibrium statistical physics specifically the replica method. Macroscopic properties of the inference problem are introduced and replica symmetric order parameter equations linking them are derived. The inclusion of a non-trivial covariance structure brings these methods closer to real-world applications.

The resulting high-dimensional asymptotic theory is first applied to the linear regression model where classical results are recovered for $p \ll N$ and the corresponding results for $p \sim N$ are derived. Next, application to the logistic regression model highlights inference with imbalanced class sizes where the literature contains many practical methods but is generally lacking theoretical results. Our replica formalism is also flexible enough to deal with time-to-event analysis in the form of the regularized Cox proportional hazards model [71, 145]. The complications of an unspecified hazard function is dealt with through a variational approximation which results in additional order parameter equations. For all three

models considered, verification of the model results is made through comparisons to existing analytical results, numerical simulations and equivalent results using other methods.

Part II: Integrable Bayesian Inference in High Dimensions. In the classical $N \to \infty$ limit, taking point estimates of unknown parameters by maximum likelihood methods is sufficient. On the other hand, for data where $p \sim \mathscr{O}(N)$ these methods are prone to overfitting [91]. In the Bayesian approach, we integrate over a distribution of model parameters by utilizing the full prior probability density. By avoiding parameter point estimation and delaying the hyperparameter point estimation until the very end of the calculation, we hope to reduce the model overfitting, potentially allowing for high-dimensional datasets to be classified.

We can proceed using numerical or analytical methods. Despite all the advantages of the former, approximating the posterior distribution numerically becomes problematic when considering a large space of models. In addition, iterations becoming stuck in local modes require heuristic solutions. Before the wide-spread usage of MCMC methods [104], Bayesian analysis was restricted to simple models and their conjugate priors. By resurrecting and further elaborating methods used before this, we attempt to widen the family of analytical priors. This approach may still be preferable to attempting to sample from a very high-dimensional posterior distribution using MCMC sampling.

Our approach proceeds by extending the multivariate Gaussian generative classifier using a generalization of the conjugate, normal-Wishart prior distribution. This allows us to derive a closed form expression for the predictive probabilities in two special cases and avoids the need for numerical approximations which are time-consuming, computational-intensive and often contribute to overfitting. In the process, we suggest alternative methods for estimating hyperparameters which are less computationally expensive than the standard cross-validation approach. The result is a novel generative Bayesian classifier which is applied to a medical data study of Terahertz frequency electromagnetic radiation where $p \sim N$.

Part I

Statistical Physics of High-Dimensional Inference

Chapter 2

Introducing the overfitting framework - Linear Regression

2.1 Statistical physics of high-dimensional inference

Classical statistical inference performs well when the ratio $p/N = \zeta \approx 0$ assuming fixed p and a diverging number of samples $N \to \infty$. However problems are encountered in the regime $\zeta > 0$. Possible methods of mitigation such as regularization and variable selection were introduced in the previous chapter. To systematically deal with this high-dimensional data regime, we look to alternative methods.

At the core of many inference problems lies an optimization task. For example, ML estimation maximizes the likelihood or log likelihood function. Mapping this optimization to a statistical physics problem has led to important advances in our understanding of statistical inference problems (see [86, 97] for an overview). By introducing an energy function, the associated degrees of freedom and the Gibbs-Boltzmann probability measure, we can utilize the tools of equilibrium statistical mechanics which was developed to calculate the bulk properties of a system with many interacting microscopic particles.

We introduce key ingredients of the minimization through the canonical example of an Ising model with state configuration σ and inverse temperature[1] $\gamma = 1/T$. The probability distribution and partition function Z are

$$p(\sigma) = \frac{1}{\mathrm{Z}} e^{-\gamma \mathscr{H}(\sigma)}, \qquad \mathrm{Z} = \sum_{\sigma} e^{-\gamma \mathscr{H}(\sigma)} \tag{2.1}$$

[1] The usual symbol β will be used to represent the regression coefficients of the model

At low temperature ($\gamma \to \infty$), the probability distribution is dominated by the minimum energy states of the system. These correspond to the required solutions of our original optimization problem. At high temperature ($\gamma \to 0$), the probability distribution is uniform with each configuration having probability $1/\mathrm{Z}$. The partition function Z plays a special role in statistical physics. It is the sum over all possible states of the system, normalizes the probability distribution and is typically computationally intractable. We will encounter Z again in the Bayesian formulation of Part II where it is known as the evidence. The equilibrium state is now found by minimizing the free energy density $f = -\frac{1}{\beta N}\log Z(\beta)$. In the thermodynamic limit $N \to \infty$, this intensive thermodynamic potential is assumed to be self-averaging i.e. independent of the particular realization of the data (the quenched disorder). Equilibrium is meant in the thermodynamic sense of the minimum free energy $F = E - TS$. The dynamics of arriving at this state, described by non-equilibrium statistical mechanics, are not considered in this thesis.

Mathematical models of physical systems often require simplification before analysis is possible e.g. Ising model with binary spins or social balance theory [20, 68] with ± 1 interactions. This is also true for models of statistical inference. In the following work, the inferred regression coefficients will be our degrees of freedom and increasing levels of complexity are possible: $\beta \in \{0,1\}^p$ (Ising model), $\sum_{\mu=1}^{p} \beta_\mu^2 = p$ (spherical model), $\beta \in \{1,2,\ldots,k\}^p$ (Potts model) and $\beta \in \mathbb{R}^p$ (continuous configuration space). See [9] for pedagogical examples of these models. Here we permit the regression coefficients to take any value on the real line and the resulting integrals stay finite due to the presence of a quadratic potential in the form of L2 regularization.

For most non-trivial systems, there exist values of γ where the free energy density is non-analytic. These points are called phase transitions and are exactly the points at which numerical simulations slow down. Hence a theoretical understanding of the location of phase transitions in the statistics mechanics problem can lead to insights into difficult areas of the corresponding inference problem. Simplifying models may have the effect of removing phase transitions e.g. mean-field approximations. This is not true for the three models considered in this thesis.

2.2 Motivation for the inference problem

We start by introducing a family of models commonly used in inference problems. These Generalized Linear Models (GLMs) [95] are characterized by the function linking the predictor $\beta \cdot \mathbf{z}$ to the response variable t. For example, in normal linear regression, this link function is $g(x) = x$. Non-linear examples include $g(x) = \log\,[x/(1-x)]$ for logistic

regression and $g(x) = \Phi^{-1}(x)$ for the probit model (where $\Phi(x)$ is the cumulative distribution function for the Gaussian distribution). To gain some intuition for the inference problem, we first consider data generated from the simplest GLM

$$t_i = r^0 + \beta^0 \cdot \mathbf{z}_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0, (\sigma^0)^2) \tag{2.2}$$

The data $\mathscr{D} = \{(t_i, \mathbf{z}_i)\}_{i=1}^N$ has the response variable $t_i \in \mathbb{R}$ and covariates $\mathbf{z}_i \in \mathbb{R}^p$. The true model parameters are the regression coefficients $\beta^0 = \{\beta_1^0, \ldots, \beta_p^0\} \in \mathbb{R}^p$ and the intercept $r^0 \in \mathbb{R}$. Finally $\varepsilon_i \sim \mathrm{N}(0, (\sigma^0)^2)$ represents Gaussian noise with variance $(\sigma^0)^2 > 0$. The intercept is intentionally treated separately from the p regression coefficients. This will become important in the analysis of imbalanced class sizes in Section 3.3.4.

Given a dataset generated by (2.2), we must assume a model in order to carry out the inference procedure. Here we stress the key working assumption of our theory that we know the true form of the data-generating model. In the normal linear model assumed in this chapter, the residuals are by definition normally distributed with the following (univariate) conditional probability.

$$p(t_i|\beta, \mathbf{z}_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(t_i - \beta\cdot\mathbf{z}_i - r)^2} \tag{2.3a}$$

$$\log p(t_i|\beta, \mathbf{z}_i) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(t_i - \beta\cdot\mathbf{z}_i - r)^2 \tag{2.3b}$$

Non-normal residuals may point to undesirable causes such as outliers in the data, covariates not included in the model or a non-linear relationship with an existing covariate (see [35] for a comprehensive discussion).

To gain some intuition, we generate data from (2.2) with $\mathbf{z}_i \sim \mathrm{N}(0,1)$, $\beta_\mu^0 \sim \mathrm{N}(0,1)$, $r^0 = 0$ and infer the regression coefficients using the Nelder-Mead simplex method [102]. Figure 2.1 plots the inferred regression coefficients $\hat{\beta}_\mu$ against the true ones β_μ^0 for $\mu = \{1, 2, \ldots, p\}$. There are no surprises in this plot: The slope is one in both cases reflecting perfect inference i.e. $\mathbb{E}(\hat{\beta}) = \beta^0$ where the expectation is over the noise distribution. Inference is possible when $p \sim N$ but the variance of the data cloud increases as the number of samples N decreases.

Equation (2.3a) can be written in multivariate form by defining $\beta = (\beta_1, \ldots, \beta_p)$, $\mathbf{t}^T = (t_1, \ldots, t_N)$ i.e. an $N \times 1$ vector of response variables, $\mathbf{z}_i^T = (z_{i1}, \ldots, z_{ip})$ forming a data matrix:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_N^T \end{pmatrix} = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ldots & \vdots \\ z_{N1} & \cdots & z_{Np} \end{pmatrix} \in \mathbb{R}^{N\times p} \tag{2.4}$$

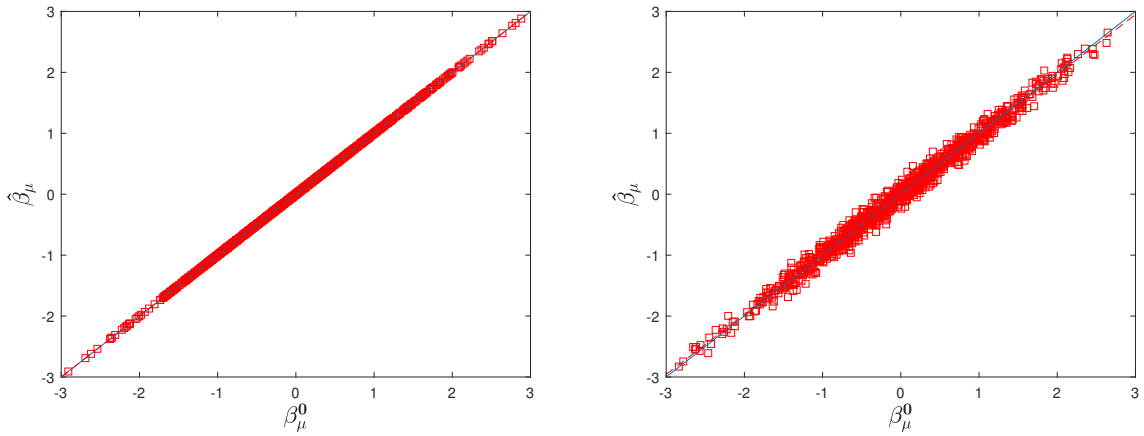

Fig. 2.1 Synthetic data is generated using model (2.2) with the true regression coefficients distributed as β^0_μ, $\varepsilon_i \sim \mathrm{N}(0,1)$, $r^0 = 0$ and data dimension, $p = 1000$. The plots show inferred $\hat{\beta}$ versus true β^0 regression coefficients for $N = 5000$ (left) and $N = 1000$ (right) samples. The slope for both data clouds is approximately one.

Re-writing (2.3b) in multivariate form, omitting the intercept for clarity and defining a $N \times N$ covariance matrix Σ

$$\log p(\mathbf{t}|\beta,\mathbf{Z}) = -\frac{p}{2}\log 2\pi - \frac{1}{2}\log \mathrm{Det}\Sigma - \frac{1}{2}(\mathbf{t}-\mathbf{Z}\beta)^T\Sigma^{-1}(\mathbf{t}-\mathbf{Z}\beta) \tag{2.5}$$

We derive classical results for the mean and variance of inferred regression coefficients for later comparison with our theory. Maximizing the log likelihood with respect to β is equivalent to minimizing the final term or error function in (2.5).

$$\nabla_\beta\left[-\frac{1}{2}(\mathbf{t}-\mathbf{Z}\beta)^T\Sigma^{-1}(\mathbf{t}-\mathbf{Z}\beta)\right] = \mathbf{Z}^T\Sigma^{-1}(\mathbf{t}-\mathbf{Z}\beta) \tag{2.6}$$

Setting this derivative equal to zero and assuming $\Sigma = \mathbb{I}$ leads to an expression for the inferred regression coefficients in terms of the data

$$\hat{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{t} \tag{2.7}$$

This expression is valid for $p \leq N$ ensuring $\mathbf{Z}^T\mathbf{Z}$ is non-singular. We will see that other GLMs do not admit a closed form solution for the inferred regression coefficients. The expected value of the inferred regression coefficients over the noise distribution confirms that $\hat{\beta}$ is an unbiased estimator agreeing with Figure 2.1

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}\big[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{t}\big] = \mathbb{E}\big[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Z}\beta^0+\varepsilon)\big] = \beta^0 \tag{2.8}$$

The covariance matrix of $\hat{\beta}$ for $p < N$ can be calculated [37] using $\mathbb{E}(\hat{\beta}) = \beta^0$

$$\begin{aligned}\mathbb{E}\big[(\hat{\beta}-\beta^0)(\hat{\beta}-\beta^0)^T\big] &= \big[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\big]\mathbb{E}\big[(\mathbf{t}-\mathbf{Z}\beta^0)(\mathbf{t}-\mathbf{Z}\beta^0)^T\big]\big[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\big]^T \\ &= \big[(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\big]\ \mathrm{Var}(\mathbf{t})\ \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1} = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}\end{aligned} \tag{2.9}$$

where we have used $\mathrm{Var}(\mathbf{t}) = \sigma^2\mathbb{I}$ and $\hat{\beta}-\beta^0 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{t}-\beta^0 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{t}-\mathbf{Z}\beta^0)$. The variance of association parameters can be estimated from the diagonal entries of $\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$

$$\begin{aligned}\mathbb{E}(\hat{\beta}_\mu^2)-[\mathbb{E}(\hat{\beta}_\mu)]^2 &= \sigma^2\big[(\mathbf{Z}^T\mathbf{Z})^{-1}\big]_{\mu\mu} \\ \mathbb{E}(\hat{\beta}_\mu\hat{\beta}_\nu)-\mathbb{E}(\hat{\beta}_\mu)\mathbb{E}(\hat{\beta}_\nu) &= \sigma^2\big[(\mathbf{Z}^T\mathbf{Z})^{-1}\big]_{\mu\nu}\end{aligned} \tag{2.10}$$

So far only uncorrelated covariates have been considered. Correlated data is typically addressed in classical statistics using Variance Inflation Factors (VIF) [93] defined for covariate μ as

$$\mathrm{VIF}_\mu = \frac{1}{1-\mathrm{R}_\mu^2} \in [1,\infty) \tag{2.11}$$

The coefficient of determination R_μ^2 is found by regressing covariate μ against the remaining data. The lower bound of VIF_μ is reached when vector $\mathbf{z}_\mu$ is orthogonal to all other covariates vectors in the data matrix $\mathbf{Z}$. It is greater than unity when this is not the case. We use this term when comparing our theoretical variance expression to classical results (2.75). The situation for non-linear GLMs is not so straightforward [92].

Having reminded ourselves of the classical results in the regime $p < N$, the inference problem is now analyzed by a direct approach in Section 2.3 and subsequently by developing a statistical physics framework in Section 2.4 in the spirit of [26]. The results are shown to agree with classical results in Section 2.5. We intentionally introduce the statistical physics formalism with a simple model before tackling progressively more complex GLMs in later chapters. We hope this leads to a pedagogical format.

2.3 Replica-free approach

Two key observables of the linear regression model are identified as the overlap between true and inferred parameters $\beta^0\cdot\langle\beta\rangle$ and the variance of the inferred parameters $\langle\beta_\mu\beta_\nu\rangle - \langle\beta_\mu\rangle\langle\beta_\nu\rangle$ where $\langle\ldots\rangle$ denotes the average over the Gibbs-Boltzmann distribution. The assumption of Gaussian noise, which is central to the linear regression model, along with the introduction of generating fields permits a direct calculation of the required expressions.

From (2.3b) and a constant noise variance σ^2, maximizing the log likelihood function with *L2* regularizer parametrized by η is equivalent to minimizing

$$\mathrm{E} = \frac{1}{2}\sum_{i=1}^{N}(t_i - \beta\cdot\mathbf{z}_i)^2 + \frac{1}{2}\eta\beta\cdot\beta \tag{2.12}$$

where the intercept has been neglected for clarity. Let $\mathbf{Z}$ be the $N\times p$ data matrix of (2.4) and define the $p\times p$ covariance matrix $\mathbf{C} = N^{-1}\mathbf{Z}^T\mathbf{Z}$. Introducing a scalar generating field λ, the required overlap function is $\langle\beta^0\cdot\beta\rangle = \frac{1}{\mathrm{Z}}\int \mathrm{d}\beta\ p(\beta)\ \beta^0\cdot\beta$

$$\begin{aligned}
\langle\beta^0\cdot\beta\rangle &= \frac{1}{\mathrm{Z}}\int \mathrm{d}\beta\ \beta^0\cdot\beta\ e^{-\frac{1}{2}\gamma\ \sum_{i=1}^{N}(t_i-\beta\cdot\mathbf{z}_i)^2-\frac{1}{2}\gamma\ \eta\beta\cdot\beta}\\
&= \lim_{\lambda\to 0}\ \frac{\partial}{\partial\lambda}\log\int \mathrm{d}\beta\ e^{\lambda\beta^0\cdot\beta-\frac{1}{2}\gamma\ \sum_{i=1}^{N}(t_i-\beta\cdot\mathbf{z}_i)^2-\frac{1}{2}\gamma\eta\beta\cdot\beta}\\
&= \lim_{\lambda\to 0}\ \frac{\partial}{\partial\lambda}\log\left\{e^{-\frac{1}{2}\gamma\ \sum_{i=1}^{N}t_i^2}\int \mathrm{d}\beta\ e^{-\frac{1}{2}\beta\cdot\gamma(N\mathbf{C}+\eta\,\mathbb{I})\beta}\ e^{\beta\cdot(\lambda\beta^0+\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i)}\right\}\\
&= \lim_{\lambda\to 0}\ \frac{\partial}{\partial\lambda}\log\frac{(2\pi)^{\frac{p}{2}}}{|\gamma(N\mathbf{C}+\eta\,\mathbb{I})|^{\frac{1}{2}}}\left\langle e^{\beta\cdot(\lambda\beta^0+\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i)}\right\rangle
\end{aligned} \tag{2.13}$$

where the final average is over the zero mean multivariate Gaussian with covariance matrix $[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}$. The regularizer $\eta > 0$ prevents this matrix inverse from being singular i.e. when $\mathrm{rank}(\mathbf{C}) < p$. Using the moment generating function $\mathbb{E}(e^{\mathbf{s}.\mathbf{x}}) = e^{\frac{1}{2}\mathbf{s}.\Sigma\mathbf{s}}$ with $\mathbf{s} = \lambda\beta^0+\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i$ (see Appendix A.1) gives

$$\begin{aligned}
\langle\beta^0\cdot\beta\rangle &= \lim_{\lambda\to 0}\ \frac{\partial}{\partial\lambda}\log e^{\frac{1}{2}(\lambda\beta^0+\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i)\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}(\lambda\beta^0+\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i)}\\
&= \lim_{\lambda\to 0}\ \frac{\partial}{\partial\lambda}\left\{\frac{1}{2}\lambda^2\beta^0\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}\beta^0+\lambda\beta^0\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}\gamma\ \sum_{i=1}^{N}t_i\mathbf{z}_i+\text{constant}\right\}\\
&= \lim_{\lambda\to 0}\left\{\lambda\beta^0\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}\beta^0+\beta^0\cdot(N\mathbf{C}+\eta\,\mathbb{I})^{-1}\ \sum_{i=1}^{N}t_i\mathbf{z}_i\right\}\\
&= \beta^0\cdot(N\mathbf{C}+\eta\,\mathbb{I})^{-1}\ \sum_{i=1}^{N}t_i\mathbf{z}_i
\end{aligned} \tag{2.14}$$

Note we have not taken the limit $\gamma\to\infty$ but our estimate for $\beta^0\cdot\beta$ for dataset $\mathscr{D} = \{(t_1,\mathbf{z}_1),\ldots,(t_N,\mathbf{z}_N)\}$ is independent of the inverse temperature γ. The vector $(N\mathbf{C}+\eta\,\mathbb{I})^{-1}\ \sum_{i=1}^{N}t_i\mathbf{z}_i$ corresponds to the $\hat{\beta}$ expression in (2.7). To compare to the result derived using the replica method in Section 2.4, we average over the true data distribution

$p(\mathbf{z},t|\beta^0)$

$$\begin{aligned}
\langle\langle\beta^0\cdot\beta\rangle\rangle_{\mathscr{D}} &= \left\langle \beta^0\cdot(N\mathbf{C}+\eta\,\mathbb{I})^{-1}\ \sum_{i=1}^{N} t_i\mathbf{z}_i \right\rangle_{\mathscr{D}} \\
&= \beta^0\cdot\int \mathrm{d}\mathbf{z}\ p(\mathbf{z})\int \mathrm{d}t\ p(t|\mathbf{z},\beta^0)\left(N\mathbf{C}+\eta\,\mathbb{I}\right)^{-1} t\ \mathbf{z} \\
&= \beta^0\cdot\int \mathrm{d}\mathbf{z}\ p(\mathbf{z})\left(N\mathbf{C}+\eta\,\mathbb{I}\right)^{-1}\mathbf{z}\int \mathrm{d}t\ t\ \mathrm{N}(t|\beta^0\cdot\mathbf{z},\ \sigma^2) \\
&= \beta^0\cdot\int \mathrm{d}\mathbf{z}\ p(\mathbf{z})\left(\mathbf{Z}^T\mathbf{Z}+\eta\,\mathbb{I}\right)^{-1}\left(\beta^0\mathbf{Z}^T\mathbf{Z}\right) \\
&= \beta^0\cdot\beta^0\int \mathrm{d}\mathbf{z}\ p(\mathbf{z})\ \frac{\mathbf{Z}^T\mathbf{Z}}{\mathbf{Z}^T\mathbf{Z}+\eta\,\mathbb{I}}
\end{aligned} \tag{2.15}$$

since $\int \mathrm{d}t\ t\ \mathrm{N}(t|\beta^0\cdot\mathbf{z},\ \sigma^2)=\beta^0\cdot\mathbf{z}$. The maximum likelihood case $\eta=0$ implies $\langle\langle\beta^0\cdot\beta\rangle\rangle_{\mathscr{D}}=\beta^0\cdot\beta^0$ ie the slope of the graph in Figure 2.1 is independent of p/N and the ML estimator is unbiased $\mathbb{E}(\hat{\beta})=\beta^0$. Expanding in small η shows the slope decreases with increasing η.

The variance of the regression coefficients can be calculated in a similar way but with a vector generating field $\lambda\in\mathbb{R}^p$ where λ_j is its j^{th} component.

$$\begin{aligned}
\langle\beta_\mu\beta_\nu\rangle-\langle\beta_\mu\rangle\langle\beta_\nu\rangle &= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\log\int \mathrm{d}\beta\ e^{\lambda\cdot\beta-\frac{1}{2}\gamma\sum_{i=1}^{N}(t_i-\beta\cdot\mathbf{z}_i)^2-\frac{1}{2}\gamma\eta\beta\cdot\beta} \\
&= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\log\int \mathrm{d}\beta\ e^{-\frac{1}{2}\beta\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]\beta}\ e^{\beta\cdot(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)} \\
&= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\log\frac{(2\pi)^{\frac{p}{2}}}{|\gamma(N\mathbf{C}+\eta\,\mathbb{I})|^{\frac{1}{2}}}\left\langle e^{\beta\cdot(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)}\right\rangle \\
&= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\log e^{\frac{1}{2}(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)} \\
&= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\left\{\frac{1}{2}(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}(\lambda+\gamma\sum_{i=1}^{N}t_i\mathbf{z}_i)\right\} \\
&= \lim_{\lambda\to 0}\ \frac{\partial^2}{\partial\lambda_\mu\partial\lambda_\nu}\left\{\frac{1}{2}\lambda\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}\ \lambda+\lambda\cdot[\gamma(N\mathbf{C}+\eta\,\mathbb{I})]^{-1}\ \gamma\sum_{i=1}^{N}t_i\mathbf{z}_i\right\} \\
&= \left([\gamma(N\mathbf{C}+\eta\mathbb{I}_p)]^{-1}\right)_{\mu\nu} = \frac{1}{\gamma}\left[\left(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^T+\eta\,\mathbb{I}\right)^{-1}\right]_{\mu\nu}
\end{aligned} \tag{2.16}$$

Note this expression is a function of the covariate data $\{\mathbf{z}_i\}_{i=1}^N$ but not the response variables $\{t_i\}_{i=1}^N$. It is identical to (2.10) in the ML case $\eta = 0$ and the inverse temperature $\gamma = 1/\sigma^2$. By calculating the derivatives of ℓ, the log likelihood (2.3b)

$$\frac{\partial \ell}{\partial \beta_\mu} = \sum_{i=1}^N (t_i - \beta \cdot \mathbf{z}_i)\, z_{i\mu} \quad \text{and} \quad \frac{\partial^2 \ell}{\partial \beta_\mu \beta_\nu} = -\sum_{i=1}^N z_{i\mu} z_{i\nu} \tag{2.17}$$

and using a definition of the Fisher information matrix $\mathrm{I}(\beta)$

$$[\mathrm{I}(\beta)]_{\mu\nu} = -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta_\mu \beta_\nu}\right) = \mathbb{E}\left(\sum_{i=1}^N \mathbf{z}_i \mathbf{z}_i^T\right)_{\mu\nu} \tag{2.18}$$

From (2.16) and (2.18), we find that the unbiased ML estimator for the normal linear regression model achieves the Cramer-Rao lower bound $[\mathrm{I}(\beta)]^{-1}$ as expected.

Although this direct calculation was successful for normal linear regression, we proceed with the more general approach using replica theory in the following section. This will provide the framework for analyzing non-linear models where the direct approach is not possible. In addition, comparison with known statistical results will provide initial validation for our more complex replica approach.

2.4 Replica analysis

2.4.1 Model setup

Assume data $\mathscr{D} = \{(t_1, \mathbf{z}_1), \ldots, (t_N, \mathbf{z}_N)\}$ where $\mathbf{z}_i \in \mathbb{R}^p$ with response variable $t_i \in \mathbb{R}$. Following [26], we choose as our measure of overfitting

$$\mathrm{E}(\vartheta, \mathscr{D}) \equiv D(\hat{P}_{\mathscr{D}} \| P_\vartheta) - D(\hat{P}_{\mathscr{D}} \| P_{\vartheta^0}) \tag{2.19}$$

where the empirical distribution of covariates and response variables is

$$\hat{P}_{\mathscr{D}} = \hat{P}(t, \mathbf{z} | \mathscr{D}) = \frac{1}{N} \sum_{i=1}^N \delta(t - t_i)\, \delta(\mathbf{z} - \mathbf{z}_i) \tag{2.20}$$

We note that (2.19) describes a shifted version of the maximum likelihood function since the true (fixed) model parameters ϑ^0 do not participate in the optimization over the parameter space. We will show this algebraically after the application of the tools of equilibrium statistical physics in (2.25).

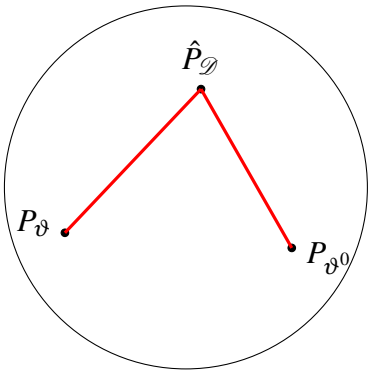

Fig. 2.2 Schematic diagram of overfitting measure E

Justification for the choice of overfitting measure. The expression (2.19) is not the only way to characterize overfitting. Three reasons for our current choice are described here.

- Minimizing the Kullback-Leibler divergence $D(\hat{P}_{\mathscr{D}}\|P_{\vartheta})$ is equivalent to maximizing the likelihood function. This provides the link to conventional statistical methods.
- Our choice provides a convenient threshold for the presence of overfitting where $\mathrm{E} > 0$ represents underfitting and $\mathrm{E} < 0$ represents overfitting. From (2.19), $\theta^0 = \hat{\theta} \Rightarrow \mathrm{E} = 0$ however we find in Section 3.1.1 that $\mathrm{E} = 0 \not\Rightarrow \theta^0 = \hat{\theta}$.
- Expression (2.19) is a function of all model parameters including the hazard rate for survival analysis. This is not always the case for other measures as shown below.

Alternative choices. There are many attempts to characterize the phenomenon of overfitting in the literature. Examples are included here are comparison with our approach. It would be interesting to attempt our analysis using these alternative formulations.

- The Akaike Information Criterion (AIC) [3] uses information theory to balance goodness-of-model-fit with model complexity i.e. model selection. The AIC is a function of the number of parameters p and the maximum likelihood value L_{ML}.

$$\mathrm{AIC} = -2\log(\mathrm{L_{ML}}) + 2\mathrm{p} \tag{2.21}$$

  When one of the model parameters is a function, as in the Cox model of Chapter 4, the number of parameters diverge so the AIC cannot be used. Other information criteria, such BIC, also involve the number of parameters limiting their use. In addition, we will see in this chapter that the MLE itself biased for $p = \mathscr{O}(N)$.

- To account for the regime where both N and p diverge, the Generalized Information Criterion (GIC) [45] can be used but at the expense of introducing an additional tuning parameter.

- An alternative to the KL divergence is the Lévy distance which measures convergence of probability distributions. This has been used, for example, to compare the true eigenvalue distribution to a derived estimator [42].

- In context of the penalized Cox proportional hazards model, the mean absolute proportion bias (MPB) [71] of the estimator was introduced

$$\mathrm{MPB}(\hat{\beta}) = \frac{1}{\mathrm{p}} \sum_{\mathrm{j}=1}^{\mathrm{p}} \left| \frac{\hat{\beta}_{\mathrm{j}}}{\beta_{\mathrm{j}}^{0}} - 1 \right| \tag{2.22}$$

  We find that in Chapter 4 that both the regression coefficients and the inferred hazard rate become biased as the ratio p/n increases. It is therefore important to include both sets of model parameters in whichever measure is chosen.

Simulations designed to gain intuition for our chosen measure are set out in Section 3.1.1. We proceed by minimizing the energy function $\mathrm{E}(\vartheta, \mathscr{D})$ by setting up an equivalent equilibrium statistical mechanics problem. The regression coefficients $\{\beta_\mu\}_{\mu=1}^{p}$ will represent the degrees of freedom and each realization of $\mathscr{D}$ is the quenched disorder. To extend the theory from maximum likelihood to maximum a posteriori estimators, a penalty term is added i.e. $D(\hat{P}_{\mathscr{D}}|P_\vartheta) - \log p(\vartheta)$. Commonly used priors when $\beta \subseteq \vartheta$ are $p(\beta) \propto \exp[-\eta \sum_{\mu=1}^{p} |\beta_\mu|]$ (giving $L1$ regularization[2], or 'LASSO' regression [136]) and $p(\beta) \propto \exp[-\eta \sum_{\mu=1}^{p} \beta_\mu^2]$ (giving $L2$ regularization, or 'ridge' regression). For $\eta > 0$, this is equivalent to imposing an (unnormalized) Gaussian prior on the likelihood. We can now write a Hamiltonian for our

[2]This choice promotes sparsity in the regression coefficient vector β, which would result in a horizontal line segment passing through the origin in Figure 2.1. Since our theory aims to predict the slope of the data clouds in Figure 2.1, we will not pursue $L1$ regularizers in this paper.

equivalent statistical physics problem where Θ is the valid parameter space

$$
\begin{aligned}
H(\vartheta|\vartheta^0,\mathscr{D}) &\equiv \min_{\vartheta\in\Theta}\left\{D(\hat{P}_{\mathscr{D}}\|P_\vartheta)-\log p(\vartheta)\right\}-\left\{D(\hat{P}_{\mathscr{D}}\|P_{\vartheta^0})-\log p(\vartheta^0)\right\}\\
&= \min_{\vartheta}\left\{\log\frac{p(\vartheta^0)}{p(\vartheta)}+\sum_{i=1}^{N}\left[\hat{P}_{\mathscr{D}}(t_i|\mathbf{z}_i)\log P_{\vartheta^0}(t_i|\mathbf{z}_i)-\hat{P}_{\mathscr{D}}(t_i|\mathbf{z}_i)\log P_\vartheta(t_i|\mathbf{z}_i)\right]\right\}\\
&= \min_{\vartheta}\left\{\log\frac{p(\vartheta^0)}{p(\vartheta)}+\sum_{i=1}^{N}\left[\frac{1}{N}\sum_{j=1}^{N}\delta(t_i-t_j)\log\frac{P_{\vartheta^0}(t_i|\mathbf{z}_i)}{P_\vartheta(t_i|\mathbf{z}_i)}\right]\right\}\\
&= \min_{\vartheta}\frac{1}{N}\left\{N\log\frac{p(\vartheta^0)}{p(\vartheta)}+\sum_{j=1}^{N}\sum_{i=1}^{N}\delta(t_i-t_j)\log\frac{P_{\vartheta^0}(t_i|\mathbf{z}_i)}{P_\vartheta(t_i|\mathbf{z}_i)}\right\}\\
&= \min_{\vartheta}\frac{1}{N}\left\{N\log\frac{p(\vartheta^0)}{p(\vartheta)}+\sum_{j=1}^{N}\log\frac{P_{\vartheta^0}(t_j|\mathbf{z}_j)}{P_\vartheta(t_j|\mathbf{z}_j)}\right\}\\
&= \min_{\vartheta}\left\{\frac{1}{N}\sum_{i=1}^{N}\log\frac{p(t_i|\mathbf{z}_i,\vartheta^0)\,p(\vartheta^0)}{p(t_i|\mathbf{z}_i,\vartheta)\;p(\vartheta)}\right\}
\end{aligned}
\tag{2.23}
$$

In the presence of an *L*2 regularizer, correlations between covariates can no longer be transformed away (see appendix B.1) leading to a more complex theory than [26] and ultimately to a mild restriction on the eigenvalue spectrum of covariance matrix. This introduces additional mathematical complications but allows for investigation of the previously inaccessible regime $\zeta > 1$. MAP regression is equivalent to minimizing the quantity (2.23) and can proceed by a number of routes:

1. Numerical optimization e.g. gradient descent methods.
2. Variational approximations. Choosing a suitable approximation for the probability distribution which factorizes.
3. Cavity method [96] allowing the investigation of macroscopic properties of a single realization of the data.
4. Replica method which produces results averaged over the data (quenched disorder).

Typically statistical physics problems rely on the assumption that the macroscopic properties of the system are not dependent on the particular realization of the disorder. The data takes this role in our inference problem and we proceed by averaging over the quenched disorder, represented by $\langle\ldots\rangle_{\mathscr{D}}$, to find the typical behaviour of our measure E. The associated

free energy expression becomes[3]

$$\begin{aligned}
\mathrm{E}_\gamma(\vartheta^0, \mathscr{D}) &= -\frac{1}{N}\frac{\partial}{\partial\gamma}\log\int d\vartheta\; e^{-\gamma\sum_{i=1}^N \log\frac{p(t_i|\mathbf{z}_i,\vartheta^0)p(\vartheta^0)}{p(t_i|\mathbf{z}_i,\vartheta)p(\vartheta)}} \\
\Rightarrow \mathrm{E}_\gamma(\vartheta^0) &= -\frac{1}{N}\frac{\partial}{\partial\gamma}\left\langle \log\int d\vartheta\; e^{-\gamma\sum_{i=1}^N \log\frac{p(t_i|\mathbf{z}_i,\vartheta^0)p(\vartheta^0)}{p(t_i|\mathbf{z}_i,\vartheta)p(\vartheta)}}\right\rangle_{\mathscr{D}} \\
\mathrm{E}_\gamma(\vartheta^0) &= -\frac{1}{N}\frac{\partial}{\partial\gamma}\left\langle \log\int \mathrm{d}\vartheta\; \prod_{i=1}^N\left[\frac{p(t_i|\mathbf{z}_i,\vartheta)\; p(\vartheta)}{p(t_i|\mathbf{z}_i,\vartheta^0)\; p(\vartheta^0)}\right]^\gamma\right\rangle_{\mathscr{D}}
\end{aligned} \tag{2.24}$$

To proceed with the average over a logarithm, we utilize the replica identity $\langle\log \mathrm{Z}\rangle = \lim_{n\to 0}\; n^{-1}\log\langle \mathrm{Z}^n\rangle$ (see appendix A.5). Hence the minimization problem of (2.23) corresponds to finding the ground state energy, $\mathrm{E}(\vartheta^0) \equiv \lim_{\gamma\to\infty}\mathrm{E}_\gamma(\vartheta^0)$ where the inverse temperature γ characterizes the measurement noise. Replica calculations in this paper follow similar lines to [123] and allow for exploration of the macroscopic properties of the inference problem. Before proceeding, we confirm that the expression involving ϑ^0 does not participate in the optimization.

$$\begin{aligned}
\mathrm{E}_\gamma(\vartheta^0, \mathscr{D}) &= -\frac{1}{N}\frac{\partial}{\partial\gamma}\log\int d\vartheta\; e^{-\gamma\sum_{i=1}^N \log\frac{p(t_i|\mathbf{z}_i,\vartheta^0)p(\vartheta^0)}{p(t_i|\mathbf{z}_i,\vartheta)p(\vartheta)}} \\
&= -\frac{1}{N}\frac{\partial}{\partial\gamma}\log\int d\vartheta\; e^{[-\gamma\sum_{i=1}^N \log p(t_i|\mathbf{z}_i,\vartheta^0)p(\vartheta^0)+\gamma\sum_{i=1}^N \log p(t_i|\mathbf{z}_i,\vartheta)p(\vartheta)]} \\
&= \frac{1}{N}\sum_{i=1}^N \log p(t_i|\mathbf{z}_i,\vartheta^0)p(\vartheta^0) - \frac{1}{N}\frac{\partial}{\partial\gamma}\log\int d\vartheta\; e^{\gamma\sum_{i=1}^N \log p(t_i|\mathbf{z}_i,\vartheta)p(\vartheta)}
\end{aligned} \tag{2.25}$$

Equation (2.24) is a general expression applicable to any parametric model with a single linear predictor. In the next section, we specify the normal linear regression model. By assuming a simple model first allows many expressions to be calculated analytically, a luxury which is not possible for more complex models. Inserting the conditional probability (2.2) with $\vartheta = \{\beta, r\}$ into (2.24) and using the replica identity

$$\begin{aligned}
\mathrm{E}_\gamma(\beta^0, r^0) = -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}r^1\ldots\mathrm{d}r^n\;\int \mathrm{d}\beta^1\ldots\mathrm{d}\beta^n\;\left\{\prod_{\alpha=1}^n\left[\frac{p(\beta^\alpha)}{p(\beta^0)}\right]^\gamma\right\} \\
\times\left\{\int \mathrm{d}\mathbf{z}\;\mathrm{d}t\; p(\mathbf{z})\, p(t|\mathbf{z},\beta^0,r^0)\prod_{\alpha=1}^n\left[\frac{p(t|\mathbf{z},\beta^\alpha,r^\alpha))}{p(t|\mathbf{z},\beta^0,r^0)}\right]^\gamma\right\}^N
\end{aligned} \tag{2.26}$$

[3] The expectation of an observable A can be expressed under the Gibbs-Boltzmann probability:

$$-\frac{\partial}{\partial\gamma}\log\int \mathrm{d}\vartheta e^{-\gamma A} = \frac{\int \mathrm{d}\vartheta\; A\; e^{-\gamma A}}{\int \mathrm{d}\vartheta\; e^{-\gamma A}} = \langle A\rangle$$

where $\{\beta^0, r^0\}$ represent the true regression and intercept parameters and $\{\beta^\alpha, r^\alpha\}$ the corresponding parameters inferred via MAP estimation. The index $\alpha \in \{1,\ldots,n\}$ over independent replicas of the system is generally a superscript to the parameter. To proceed with the analytical treatment, we assume the covariates have zero mean and arbitrary population covariance matrix[4]. Scaling the linear predictor, $\beta \cdot \mathbf{z} \to \beta \cdot \mathbf{z}/\sqrt{p}$ results in $\beta \cdot \mathbf{z} \sim \mathscr{O}(1)$ preventing response variables t diverging in our asymptotic theory. Replacing β^0 with β^0 to allow for more compact notation and introducing the vector $\mathbf{y}$ via the Dirac delta function to proceed with the integral:

$$p(\mathbf{y}|\beta^0,\ldots,\beta^n) = \int d\mathbf{z}\; p(\mathbf{z}) \prod_{\alpha=0}^{n} \delta\Big[y^\alpha - \frac{\beta^\alpha \cdot \mathbf{z}}{\sqrt{p}}\Big] \tag{2.27}$$

where $\mathbf{y} = \{y^0, y^1, \ldots, y^n\} \in \mathbb{R}^{n+1}$. Our energy density becomes

$$\begin{aligned} \mathrm{E}_\gamma(\beta^0, r^0) = -\frac{\partial}{\partial \gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \mathrm{d}r^1 \ldots \mathrm{d}r^n \int \mathrm{d}\beta^1 \ldots \mathrm{d}\beta^n \prod_{\alpha=1}^{n} \left[\frac{p(\beta^\alpha)}{p(\beta^0)}\right]^\gamma \\ \times \left\{ \int_{\mathbf{y}\in\mathbb{R}^{n+1}} d\mathbf{y}\; p(\mathbf{y}|\beta^0,\ldots,\beta^n) \int \mathrm{d}t\; p(t|\mathbf{y}^0, r^0) \prod_{\alpha=1}^{n} \left[\frac{p(t|\mathbf{y}^\alpha, r^\alpha)}{p(t|\mathbf{y}^0, r^0)}\right]^\gamma \right\}^N \end{aligned} \tag{2.28}$$

We assume the covariates $\{\mathbf{z}_i\}_{i=1}^N$ are independent samples from a stationary distribution $p(\mathbf{z})$ and hence so are the scalar values $\beta \cdot \mathbf{z}_1, \beta \cdot \mathbf{z}_2, \ldots, \beta \cdot \mathbf{z}_N$. In a wide range of cases described below, the central limit theorem allows us to write

$$p(\mathbf{y}|\beta^0,\ldots,\beta^n) = \frac{e^{-\frac{1}{2}\mathbf{y}^T \mathbf{C}^{-1}[\{\beta\}]\mathbf{y}}}{\sqrt{(2\pi)^{n+1} \det \mathbf{C}[\{\beta\}]}} \tag{2.29}$$

where the entries of the $(n+1)\times(n+1)$ symmetric overlap matrix $\mathbf{C}[\{\beta\}]$ are

$$\mathbf{C}_{\alpha\rho}[\{\beta\}] = \int \mathrm{d}\mathbf{z}\; \mathrm{p}(\mathbf{z}) \left(\frac{\beta^\alpha \cdot \mathbf{z}}{\sqrt{\mathrm{p}}}\right)\left(\frac{\beta^\rho \cdot \mathbf{z}}{\sqrt{\mathrm{p}}}\right) = \frac{1}{\mathrm{p}} \beta^\alpha \cdot \mathbf{A} \beta^\rho \tag{2.30}$$

and $A_{\mu\nu} = \langle z_\mu z_\nu \rangle$ defines the $p \times p$ population covariance matrix. This assumption, which is a direct consequence of working in the limit $p \to \infty$ and the Central Limit Theorem is validated numerically in Section 4.3.1.

[4]The covariates may be measured in different units i.e. height in metres and length in mm. Ridge regression would penalise large values more. Therefore we transform covariates into zero mean and variance one random variables $z = \frac{x-\mu}{\sigma}$ before doing the regression.

To show $\mathbf{C}$ is positive definite, we consider an arbitrary vector, $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^{(n+1)}$ and the notation β_i^k represents the i^{th} component of the k^{th} replica of β

$$\mathbf{x}^T\mathbf{C}\mathbf{x} = \sum_{k,l=1}^{p} x_k C_{kl} x_l = \sum_{k,l=1}^{p} x_k \left(\frac{1}{p} \sum_{i,j=1}^{p} \beta_i^k \beta_j^l A_{ij} \right) x_l = \frac{1}{p} \sum_{i,j,k,l} (x_k \beta_i^k) A_{ij} (\beta_j^l x_l) > 0 \quad (2.31)$$

since $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for positive definite covariance matrix $\mathbf{A}$.

Assumptions underlying the Central Limit Theorem. The correctly normalized sum of independent random variables typically converges (under weak conditions on the underlying distributions [46]) to a Gaussian distribution even if the original variables themselves are not normally distributed. We note that variants of the central limit theorem permit convergence to a Gaussian distribution for non-identically distributed and dependent variables under certain conditions. This is used to justify the multivariate Gaussian form of (2.29) and specifically that our covariates $\{\mathbf{z}_i\}_{i=1}^N$ need not be normally distributed. Figure 4.3 shows the macroscopic observables of our analysis do not change materially when non-Gaussian distributed covariates are used. Note that a mild restriction is later imposed on the amount of correlation between covariates due to self-averaging considerations (see Appendix B.4)

Other forms of regularization. It is possible that only a finite subset of the covariates are relevant to determine the regression outcome. This would suggest a different choice of regularization e.g. $L0$ or $L1$ known to promote sparsity on the model parameters [17, 55]. We did not pursue these alternatives here since the integrals over β (see the final line of (2.33)) would be significantly more complicated. In addition, an extensive fraction of non-zero components are required for convergence to a multivariate Gaussian distribution in (2.29).

The entries of $\mathbf{C}$ measure the similarity between the p-dimensional vectors formed by the regression parameters in different replicas. For each replica pair (α, ρ), we use the integral representation of the Dirac delta function (see Appendix A.4), and rescale the conjugate integration parameter by p to form the $(n+1) \times (n+1)$ matrix $\hat{\mathbf{C}} \equiv \{\hat{\mathbf{C}}_{\alpha\rho}\}_{\alpha,\rho=\mathbf{0}}^{\mathbf{n}}$

$$1 = \int \mathrm{d}C_{\alpha\rho}\ \delta\left[C_{\alpha\rho} - \frac{1}{p}\beta^\alpha \cdot \mathbf{A}\beta^\rho\right] = \int \frac{\mathrm{d}C_{\alpha\rho}\mathrm{d}\hat{C}_{\alpha\rho}}{2\pi/p} \mathrm{e}^{\mathrm{ip}\hat{\mathrm{C}}_{\alpha\rho}(\mathrm{C}_{\alpha\rho} - \frac{1}{\mathrm{p}}\beta^\alpha \cdot \mathbf{A}\beta^\rho)} \quad (2.32)$$

in order to simplify expression (2.28) to

$$\begin{aligned}\mathrm{E}_\gamma(\beta^0,r^0) = -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}r^1\ldots\mathrm{d}r^n \int \mathrm{d}\mathbf{C}\,\mathrm{d}\hat{\mathbf{C}}\ \frac{\mathrm{e}^{\mathrm{ip}\sum_{\alpha,\rho=0}^{\mathrm{n}}\hat{\mathrm{C}}_{\alpha\rho}\mathrm{C}_{\alpha\rho}}}{(2\pi/p)^{(n+1)^2}} \\ \times\left[\int\frac{\mathrm{d}\mathbf{y}\ \mathrm{e}^{-\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}}}{\sqrt{(2\pi)^{n+1}\det\mathbf{C}}}\int\mathrm{d}t\ p(t|\mathbf{y}^0,r^0)\prod_{\alpha=1}^{n}\left[\frac{p(t|y^\alpha,r^\alpha)}{p(t|y^0,r^0)}\right]^\gamma\right]^N \\ \times\int\mathrm{d}\beta^1\ldots d\beta^n\ \mathrm{e}^{-\eta\gamma\sum_{\alpha=1}^{\mathrm{n}}[(\beta^\alpha)^2-(\beta^0)^2]-\mathrm{i}\sum_{\alpha,\rho=0}^{\mathrm{n}}\hat{\mathrm{C}}_{\alpha\rho}\beta^\alpha\cdot\mathbf{A}\beta^\rho}\end{aligned} \tag{2.33}$$

The quadratic nature of the exponent in the β integral, a consequence of having chosen $L2$ regularization, allows for a closed form solution. Changing the penalty term to $L1$ or Lq with $q>2$ would significantly complicate the integrals.

2.4.2 Conversion into a saddle point problem

From (2.19), the overfitting measure E is a real valued function. Since the conjugate order parameter $\hat{C}$ is always associated with $\mathrm{i}=\sqrt{-1}$, it must be purely imaginary suggesting the transformation[5] $\hat{\mathbf{C}}=-\frac{1}{2}\mathrm{i}\mathbf{D}$. Before proceeding with the Laplace method of integration, we evaluate the Gaussian β integral in (2.33) by defining the $np\times np$ matrix Ξ and the np-dimensional vector ξ, with entries

$$\Xi_{\alpha\mu;\beta\nu}=2\eta\gamma\delta_{\alpha\beta}(\mathbf{A}^{-1})_{\mu\nu}+\delta_{\mu\nu}D_{\alpha\beta},\qquad \xi_\mu^\alpha=-D_{0\alpha}\tilde{\beta}_\mu^0 \tag{2.34}$$

where we have introduced the short-hand $\tilde{\beta}\equiv\mathbf{A}^{\frac{1}{2}}\beta$. With these definitions we may write the Gaussian integral in (2.33) as

$$\begin{aligned}&\int\Big(\prod_{\alpha=1}^{n}\mathrm{d}\tilde{\beta}^\alpha\mathrm{e}^{-\eta\gamma\tilde{\beta}^\alpha\cdot\mathbf{A}^{-1}\tilde{\beta}^\alpha}\Big)\mathrm{e}^{-\frac{1}{2}\sum_{\alpha,\rho=1}^{n}D_{\alpha\rho}\tilde{\beta}^\alpha\cdot\tilde{\beta}^\rho-\sum_{\rho=1}^{n}D_{0\rho}\tilde{\beta}^0\cdot\tilde{\beta}^\rho} \\ &=\mathrm{e}^{\frac{1}{2}\xi\cdot\Xi^{-1}\xi}\int\mathrm{d}\tilde{\beta}\ \mathrm{e}^{-\frac{1}{2}(\tilde{\beta}-\Xi^{-1}\xi)\cdot\Xi(\tilde{\beta}-\Xi^{-1}\xi)}=\frac{(2\pi)^{\frac{np}{2}}}{\sqrt{\det\Xi}}\mathrm{e}^{\frac{1}{2}\xi\cdot\Xi^{-1}\xi}\end{aligned} \tag{2.35}$$

Appendices B.2, B.3 contain fuller details of integral (2.35) and an expansion of the relevant terms $\xi\cdot\Xi^{-1}\xi$ and $\log\mathrm{Det}\Xi$ under the replica symmetric ansatz. Let $\{a_\mu\}$ and $\{b_\alpha\}$ denote the eigenvalues of $\mathbf{A}$ and $\mathbf{D}$, respectively. The two terms $2\eta\gamma\delta_{\alpha\beta}(\mathbf{A}^{-1})_{\mu\nu}$ and $\delta_{\mu\nu}D_{\alpha\beta}$ of the matrix Ξ commute (see (B.7)). The complete set of eigenvectors of Ξ can therefore be written as $\{\hat{\mathbf{u}}^{\mu\alpha}\}$, with components $\hat{u}^{\mu\alpha}_{\nu\rho}=u^\alpha_\rho v^\mu_\nu$, and where $\sum_{\rho\le n}D_{\lambda\rho}u^\alpha_\rho=b_\alpha u^\lambda_\rho$ and $\sum_{\nu\le p}A_{\lambda\nu}v^\mu_\nu=$

[5]Equivalently in component form $\hat{C}_{\mu\nu}=-\frac{1}{2}\mathrm{i}D_{\mu\nu}$ with the $(n+1)\times(n+1)$ matrix of conjugate order parameters $\mathbf{D}\equiv\{D_{\alpha\rho}\}_{\alpha,\rho=1}^{n}$

$a_\mu v_\lambda^\mu$, and where both are normalized according to $\sum_{\rho\leq n}(u_\rho^\alpha)^2=\sum_{\nu\leq p}(v_\nu^\mu)^2=1$. The eigenvalues of Ξ are then $\xi_{\mu\alpha}=2\eta\gamma/a_\mu+b_\alpha$, and

$$\det\Xi=\prod_{\mu=1}^{p}\prod_{\alpha=1}^{n}\left(\frac{2\eta\gamma}{a_\mu}+b_\alpha\right),\qquad(\Xi^{-1})_{\alpha\mu;\alpha'\mu'}=\sum_{\beta=1}^{n}\sum_{\nu=1}^{p}\frac{u_\alpha^\beta v_\mu^\nu u_{\alpha'}^\beta v_{\mu'}^\nu}{2\eta\gamma/a_\nu+b_\beta}\tag{2.36}$$

Hence the integral (2.35) can be written as

$$\frac{(2\pi)^{\frac{np}{2}}}{\sqrt{\det\Xi}}\,\mathrm{e}^{\frac{1}{2}\xi\cdot\Xi^{-1}\xi}=\mathrm{e}^{\frac{1}{2}np\log(2\pi)-\frac{1}{2}np\left\langle\log(2\eta\gamma/a+b)\right\rangle+\frac{1}{2}np\left\langle(\xi\cdot\hat{\mathbf{u}})^2(2\eta\gamma/a+b)^{-1}\right\rangle}\tag{2.37}$$

where the averages in the exponents are over the eigenvalues and orthonormal eigenvectors of Ξ, i.e. $\langle f(a,b,\hat{\mathbf{u}})\rangle=(np)^{-1}\sum_{\mu=1}^{p}\sum_{\alpha=1}^{n}f(a_\mu,b_\alpha,\hat{\mathbf{u}}^{\mu\alpha})$. Since $p=\zeta N$ with $\zeta>0$, the integrals over $\mathbf{C}$, $\hat{\mathbf{C}}$ and the base hazard rates in (2.33) can for $N\to\infty$ be evaluated by steepest descent, provided the limits $n\to0$ and $N\to\infty$ commute. Expression (2.37) then enables us to write the result as

$$\lim_{N\to\infty}\mathrm{E}_\gamma(\beta^0,r^0)=\frac{\partial}{\partial\gamma}\lim_{n\to0}\frac{1}{n}\mathrm{extr}\,\Psi(\mathbf{C},\mathbf{D},r^1\ldots r^n)\tag{2.38}$$

in which

$$\begin{aligned}\Psi(\mathbf{C},\mathbf{D},r^1\ldots r^n)=&-\frac{1}{2}\zeta\left[\sum_{\alpha,\rho=0}^{n}D_{\alpha\rho}C_{\alpha\rho}-\frac{1}{p}D_{00}(\tilde{\beta}^0)^2\right]+\frac{1}{2}(n+1-n\zeta)\log(2\pi)\\&+\frac{1}{2}\log\mathrm{Det}\mathbf{C}-n\eta\zeta\gamma S^2+\frac{1}{2}n\zeta\left\langle\log\left(\frac{2\eta\gamma}{a}+b\right)\right\rangle-\frac{1}{2}n\zeta\left\langle\frac{(\xi\cdot\hat{\mathbf{u}})^2}{2\eta\gamma/a+b}\right\rangle\\&-\log\int\mathrm{d}\mathbf{y}\;\mathrm{e}^{-\frac{1}{2}\mathbf{y}\cdot\mathbf{C}^{-1}\mathbf{y}}\int\mathrm{d}t\;p(t|y_0,r_0)\prod_{\alpha=1}^{n}\left[\frac{p(t|y^\alpha,r^\alpha)}{p(t|y^0,r^0)}\right]^\gamma\end{aligned}\tag{2.39}$$

where we find our theory only depends on the true regression coefficients through their variance $S^2=\lim_{p\to\infty}p^{-1}(\beta^0)^2$. Differentiating Ψ with respect to D_{00} removes D_{00} from the problem resulting in the following useful expression

$$\begin{aligned}\tilde{S}^2\equiv C_{00}&=\lim_{p\to\infty}\frac{1}{p}\beta^0\cdot\mathbf{A}\beta^0=\lim_{p\to\infty}\frac{1}{p}\sum_{\mu,\nu=1}^{p}\langle\beta_\mu^0\beta_\nu^0\rangle A_{\mu\nu}\\&=S^2\frac{1}{p}\mathrm{Tr}\mathbf{A}=S^2\int\mathrm{d}a\;\rho(a)\;a=S^2\langle a\rangle\end{aligned}\tag{2.40}$$

The other two conjugate order parameters are related to the covariance matrix $\mathbf{A}$ and regularization term η and cannot be eliminated. Details of the self-averaging argument applied to $\mathbf{A}$ are given in Appendix B.4.

2.4.3 Replica symmetric solution

To proceed, we use the replica symmetric (RS) ansatz, which assumes ergodicity of the stochastic regression process, and translates into invariance of all order parameters under all permutations of the replicas $\{1,\dots,n\}$. By thinking of the $L2$ regularization as confining the regression coefficients to a quadratic potential well, this assumption is expected to produce satisfactory results.

$$\mathbf{C}=\begin{pmatrix} C_{00} & c_0 & \dots & \dots & c_0 \\ c_0 & C & c & \dots & c \\ \vdots & c & C & \dots & c \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_0 & c & c & \dots & C \end{pmatrix} \quad \text{and} \quad \mathbf{C}^{-1}=\begin{pmatrix} B_{00} & b_0 & \dots & \dots & b_0 \\ b_0 & B & b & \dots & b \\ \vdots & b & B & \dots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_0 & b & b & \dots & B \end{pmatrix}$$

We now proceed to simplify terms in (2.39) using this RS ansatz. The eigenvalues and eigenvectors of $\mathbf{C}$, $\mathbf{C}^{-1}$ and $\mathbf{D}$ are found in [26]. $\mathbf{C}$ has two nondegenerate eigenvalues $\lambda_\pm$ with $\lambda_+\lambda_- = [C+(n-1)c]C_{00}-nc_0^2$, and a further $n-1$ fold degenerate eigenvalues $\lambda_0 = C-c$. The replica identity demands the limit $n\to 0$ so we can simplify $\log\mathrm{Det}\mathbf{C}$ by assuming small n

$$\begin{aligned}\log\mathrm{Det}\mathbf{C} &= \log\Big([C+(n-1)c]C_{00}-nc_0^2\Big)+(n-1)\log(C-c)\\ &= \log C_{00}+n\log(C-c)+\frac{n\big(c-c_0^2/C_{00}\big)}{C-c}+\mathscr{O}(n^2)\end{aligned} \tag{2.41}$$

The entries of $\mathbf{C}^{-1}$ are found by considering diagonal and off-diagonal terms of $\mathbf{C}^{-1}\mathbf{C}$

$$\begin{aligned} B_{00} &= \frac{C+(n-1)c}{C_{00}[C+(n-1)c]-nc_0^2}, & b_0 &= -\frac{c_0}{C_{00}[C+(n-1)c]-nc_0^2}\\ B &= b+\frac{1}{C-c}, & b &= \frac{c_0^2-cC_{00}}{(C_{00}[C+(n-1)c]-nc_0^2)(C-c)}\end{aligned} \tag{2.42}$$

Recalling $\mathbf{y} = \{y^0, y^1, \ldots, y^n\} \in \mathbb{R}^{n+1}$, we write an explicit expression for the quadratic form

$$\mathbf{y} \cdot \mathbf{C}^{-1}\mathbf{y} = B_{00}(y^0)^2 + (B-b)\sum_{\alpha=1}^{n}(y^\alpha)^2 + b\Big(\sum_{\alpha=1}^{n} y^\alpha\Big)^2 + 2b_0 y^0 \sum_{\alpha=1}^{n} y^\alpha \tag{2.43}$$

Next we turn to terms in (2.39) that involve the spectrum of $\mathbf{D}$. This matrix has one eigenvalue $D+(n-1)d$ with eigenvector $\mathbf{u} = (1,\ldots,1)$, and the $n-1$ fold degenerate eigenvalue $D-d$ with eigenspace $\sum_{i=1}^{n}\mathbf{u}_i = 0$. Hence

$$\begin{aligned}\left\langle \log\Big(\frac{2\eta\gamma}{a}+b\Big)\right\rangle &= \frac{1}{np}\sum_{\mu=1}^{p}\Big[(n-1)\log\Big(\frac{2\eta\gamma}{a}+D-d\Big)+\log\Big(\frac{2\eta\gamma}{a}+D-d+nd\Big)\Big] \\ &= \left\langle \log\Big(\frac{2\eta\gamma}{a}+D-d\Big)\right\rangle + \left\langle \frac{da}{2\eta\gamma+(D-d)a}\right\rangle + \mathcal{O}(n)\end{aligned} \tag{2.44}$$

Similarly, using the RS form of $\xi_\mu^\alpha = -d_0(\mathbf{A}^{\frac{1}{2}}\beta^0)_\mu$, we may write

$$\begin{aligned}\left\langle \frac{(\xi\cdot\hat{\mathbf{u}})^2}{2\eta\gamma/a+b}\right\rangle &= \frac{1}{np}\sum_{\mu=1}^{p}\Big(\sum_{\nu=1}^{p}\sum_{\rho=1}^{n}(\mathbf{A}^{\frac{1}{2}}\beta^0)_\nu v_\nu^\mu \frac{1}{\sqrt{n}}\Big)^2 \frac{d_0^2}{2\eta\gamma/a_\mu+D+(n-1)d} \\ &= \frac{1}{p}\sum_{\mu=1}^{p}(\beta^0\cdot\mathbf{v}^\mu)^2\frac{d_0^2 a_\mu}{2\eta\gamma/a_\mu+D-d}+\mathcal{O}(n) \\ &= d_0^2\left\langle \frac{a^2(\beta^0\cdot\mathbf{v})^2}{2\eta\gamma+(D-d)a}\right\rangle + \mathcal{O}(n)\end{aligned} \tag{2.45}$$

The averages in (2.44) and (2.45) are now over the joint distribution of eigenvalues and eigenvectors of $\mathbf{A}$ only. Inserting the above RS expressions into (2.39), and using $C_{00} = \tilde{S}^2$,

then gives us, with the short-hand $\mathrm{D}z = (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}z^2}\mathrm{d}z$,

$$\begin{aligned}
\frac{1}{n}\Psi_{\rm RS}(\ldots) &= -\frac{1}{2}\zeta(2d_0c_0+DC-dc)+\frac{1}{2}(1-\zeta)\log(2\pi)-\eta\zeta\gamma S^2+\mathscr{O}(n)\\
&+\frac{1}{2}\Big[\log(C-c)+\frac{c-c_0^2/C_{00}}{C-c}\Big]-\frac{1}{2}\zeta d_0^2\,\left\langle\frac{a^2(\beta^0\cdot\mathbf{v})^2}{2\eta\gamma+(D-d)a}\right\rangle\\
&+\frac{1}{2}\zeta\left\langle\log\Big(\frac{2\eta\gamma}{a}+D-d\Big)\right\rangle+\frac{1}{2}\zeta\Big\langle\frac{da}{2\eta\gamma+(D-d)a}\Big\rangle+\frac{1}{n}\log\tilde{S}\\
&-\frac{1}{n}\log\int\mathrm{D}z\int\frac{\mathrm{d}y_0}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}B_{00}y_0^2}\int\mathrm{d}t\;p(t|y_0,r^0)\\
&\times\Big[\int\mathrm{d}y\;\mathrm{e}^{-\frac{1}{2}(B-b)y^2+y(\mathrm{i}z\sqrt{b}-b_0y_0)}\frac{p^\gamma(t|y,r)}{p^\gamma(t|y_0,r^0)}\Big]^n\\
&=-\frac{1}{2}\zeta(2d_0c_0+DC-dc)+\frac{1}{2}(1-\zeta)\log(2\pi)-\eta\zeta\gamma S^2+\mathscr{O}(n)\\
&+\frac{1}{2}\Big[\log(C-c)+\frac{c-c_0^2/\tilde{S}^2}{C-c}\Big]-\frac{1}{2}\zeta d_0^2\,\left\langle\frac{a^2(\beta^0\cdot\mathbf{v})^2}{2\eta\gamma+(D-d)a}\right\rangle\\
&+\frac{1}{2}\zeta\Big\langle\log\Big(\frac{2\eta\gamma}{a}+D-d\Big)\Big\rangle+\frac{1}{2}\zeta\Big\langle\frac{da}{2\eta\gamma+(D-d)a}\Big\rangle+\frac{1}{2n}\log(\tilde{S}^2B_{00})\\
&-\frac{1}{n}\log\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\;p(t|y_0/\sqrt{B_{00}},r^0)\\
&\quad\times\Big[\int\mathrm{d}y\;\mathrm{e}^{-\frac{1}{2}(B-b)y^2+y(\mathrm{i}z\sqrt{b}-b_0y_0/\sqrt{B_{00}})}\frac{p^\gamma(t|y,r)}{p^\gamma(t|y_0/\sqrt{B_{00}},r^0)}\Big]^n
\end{aligned}\tag{2.46}$$

We note that in the replica limit $n\to 0$, (2.42) becomes

$$\begin{aligned}
&B_{00}^{-1}=\tilde{S}^2-nc_0^2/(C-c)+\mathscr{O}(n^2), && B-b=1/(C-c)\\
&b_0=-c_0/\tilde{S}^2(C-c)+\mathscr{O}(n), && b=\frac{c_0^2-c\tilde{S}^2}{\tilde{S}^2(C-c)^2}+\mathscr{O}(n)
\end{aligned}\tag{2.47}$$

Putting these results together and using the variable transformation $y\to y+\frac{wy^0+vz}{u}$, we arrive at

$$\begin{aligned}
\Psi_{RS}&=\frac{1}{2}\frac{c}{C-c}-\frac{1}{2}\zeta\log 2\pi+\frac{1}{2}\frac{\zeta c_0^2}{S^2\int\frac{\mathrm{d}a\,\rho(a)\,a^2}{2\eta\gamma+(D-d)a}}-\frac{1}{2}\zeta(CD-cd)-\eta\zeta\gamma S^2\\
&+\frac{1}{2}\zeta\int\mathrm{d}a\,\rho(a)\left[\log\left(\frac{2\eta\gamma}{a}+D-d\right)+\frac{ad}{2\eta\gamma+(D-d)a}\right]\\
&-\int\mathrm{D}z\,\mathrm{D}y_0\;\int\mathrm{d}t\;p(t|\tilde{S}y_0,r_0)\log\int Dy\;e^{y\left(\frac{y_0c_0}{\tilde{S}(C-c)}+z\sqrt{\frac{c-c_0^2/C_{00}}{(C-c)}}\right)}\Big[\frac{p(t|y\sqrt{C-c},r)}{p(t|\tilde{S}y_0,r^0)}\Big]^\gamma
\end{aligned}\tag{2.48}$$

2.4.4 Transformation of order parameters

We make the following transformations

$$\begin{aligned} u &= \sqrt{C-c}, \qquad v = \sqrt{c - c_0^2/C_{00}} = \sqrt{c - c_0^2/\tilde{S}^2} \\ w &= c_0/\sqrt{C_{00}} = c_0/\tilde{S}, \qquad f = d, \qquad g = D-d \end{aligned} \tag{2.49}$$

leading to the inverse transformations

$$c_0 = \tilde{S}w, \qquad c = v^2 + w^2, \qquad C = u^2 + v^2 + w^2 \tag{2.50}$$

where u, v, w are non-negative. These new order parameters characterize the macroscopic properties of the original inference problem and can be used to predict the behaviour of the slope and variance of the data cloud in Figure 2.1. This was achieved in the maximum likelihood case [26] through $\lim_{p\to\infty} \frac{1}{p}\beta^0 \cdot \langle\langle\beta\rangle\rangle_{\mathscr{D}}$ and $\lim_{p\to\infty} \frac{1}{p}\langle\langle\beta^2\rangle - \langle\beta\rangle^2\rangle_{\mathscr{D}}$. In our regularized case, the population covariance matrix $\mathbf{A}$ cannot be transformed away and $\beta^0 \cdot \langle\langle\beta\rangle\rangle_{\mathscr{D}}$ cannot be obtained, in general, from $\beta^0 \cdot \mathbf{A}\langle\langle\beta\rangle\rangle_{\mathscr{D}}$.

Applying the transformations and removing terms constant with respect to the six remaining order parameters $\{u, v, w, r, f, g\}$, results in a convenient finite temperature form of (2.48)

$$\begin{aligned} \Psi_{\mathrm{RS}}(\ldots) = &-\frac{1}{2}\zeta(g+f)u^2 - \frac{1}{2}\zeta g(v^2+w^2) \\ &+ \frac{1}{2}\zeta\left\{ w^2\langle a\rangle \left\langle \frac{a^2}{2\eta\gamma+ga} \right\rangle^{-1} + \left\langle \log(2\eta\gamma+ga) \right\rangle + f \left\langle \frac{a}{2\eta\gamma+ga} \right\rangle \right\} \\ &- \int \mathrm{D}z\, \mathrm{D}y_0 \int \mathrm{d}t\; p(t|\tilde{S}y_0, r_0) \log \int Dy\; p^\gamma(t|uy + wy_0 + vz, r) \end{aligned} \tag{2.51}$$

This expression can be split into a model independent Ψ^A_{RS} and the model specific term Ψ^B_{RS}:

$$\begin{aligned} \Psi^A_{\mathrm{RS}} \equiv &-\frac{1}{2}\zeta(g+f)u^2 - \frac{1}{2}\zeta g(v^2+w^2) \\ &+ \frac{1}{2}\zeta\left\{ w^2\langle a\rangle \left\langle \frac{a^2}{2\eta\gamma+ga} \right\rangle^{-1} + \left\langle \log(2\eta\gamma+ga) \right\rangle + f \left\langle \frac{a}{2\eta\gamma+ga} \right\rangle \right\} \\ \Psi^B_{\mathrm{RS}} \equiv &\int \mathrm{D}z\, \mathrm{D}y_0 \int \mathrm{d}t\; p(t|\tilde{S}y_0, r_0) \log \int Dy\; p^\gamma(t|uy + wy_0 + vz, r) \end{aligned} \tag{2.52}$$

This will be useful in subsequent chapters where different models are considered. We will only be interested in the limit $\gamma \to \infty$, where the stochastic process becomes deterministic MAP inference. This limit is taken after suitable scaling in Section 2.4.6.

2.4.5 Interpretation of order parameters

Before proceeding, we consider the correspondence between the replica symmetric order parameters and the original inference problem. Empirical evidence suggests

$$\langle \beta \rangle = \kappa \beta^0 + \omega \tag{2.53}$$

is a plausible model for the data cloud[6] in Figure 2.1. As a reminder, β^0 represents the true regression coefficients from which the data is generated. By relating this equation of a straight line to the order parameters, we find expressions for the macroscopic quantities required. The only difference from [26] is the assumption of correlated noise $\langle \omega_\mu \omega_\nu \rangle = \Omega_{\mu\nu}$. The entries $\Omega_{\mu\nu}$ of the $p \times p$ noise covariance matrix Ω have the same dimension as those of $\mathbf{A}^{-1}$, which prompt us to postulate that $\Omega = \sigma^2 \mathbf{A}^{-1}$.

Since the initial submission of this thesis, substantial progress has been made on the form of the relationship between $\langle \beta \rangle$ and β^0 for Generalized Linear Models. The calculation involves analytically evaluating the joint probability distribution $p(\beta|\beta^0)$. Here we sketch the initial stages of an involved calculation fully detailed in [27].

$$p(\beta, \beta^0 | \mathscr{D}) = \lim_{\gamma \to \infty} \frac{1}{p} \sum_{\mu=1}^{p} \frac{\int \mathrm{d}\vartheta \mathrm{d}\beta \ \mathrm{e}^{\gamma \log P(\vartheta, \beta | \mathscr{D})} \delta(\beta - \beta_\mu)}{\int \mathrm{d}\vartheta \mathrm{d}\beta \ \mathrm{e}^{\gamma \log P(\vartheta, \beta | \mathscr{D})}} \delta(\beta^0 - \beta^0_\mu) \tag{2.54}$$

with the posterior parameter likelihood

$$p(\vartheta, \beta | \mathscr{D}) = \frac{p(\beta) p(\vartheta) \prod_{i=1}^{N} p(t_i | \beta \cdot z_i / \sqrt{p}, \vartheta)}{\int \mathrm{d}\beta' \mathrm{d}\vartheta' \ p(\beta') p(\vartheta') \prod_{i=1}^{N} p(t_i | \beta' \cdot z_i / \sqrt{p}, \vartheta')} \tag{2.55}$$

The calculation can be found in the collaborative work of [27]. It uses an alternative form of the replica identity:

$$\left\langle \frac{\int \mathrm{d}x \ w(x|y) f(x)}{\int \mathrm{d}x \ w(x|y)} \right\rangle_y = \lim_{n \to 0} \left\langle \left(\int \mathrm{d}x \ w(x|y) f(x) \right) \left(\int \mathrm{d}x \ w(x|y) \right)^{n-1} \right\rangle_y \tag{2.56}$$

We replace $x \to (\vartheta, \beta)$, $y \to \mathscr{D}$, $w(x|y) \to [p(\beta) p(\vartheta) \prod_{i=1}^{N} p(t_i | \beta \cdot z_i / \sqrt{p}, \vartheta)]^\gamma$ and $f(x) \to \delta(\beta - \beta_\mu)$ and average over the data $\mathscr{D}$. By re-using the main integrals from our original

[6] The model residuals in simulations are confirmed to be normally distributed around the least squared fitted regression line

replica analysis, we find for uncorrelated covariates

$$\lim_{N\to\infty} p(\beta|\beta^0) = \frac{1}{v\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\beta - w\beta^0/S)^2/v^2} \tag{2.57}$$

We have found that the required relationship assuming uncorrelated covariates is in fact linear with Gaussian noise. This calculation is valid for any model with a single linear predictor $\beta \cdot z$. Extensive simulations across a range of models also suggest (2.53) is valid.

The correlated case, also found in [27], involves a similar calculation. Here we find that MAP regression with Gaussian priors will in the regime of finite $\zeta > 0$ not just rescale the length of the inferred association vectors but will also change its direction, irrespective of how intelligently we choose the hyperparameter η. Only for small η or weak correlations (or if by accident the vector β^0 happens to be an eigenvector of $\mathbf{A}$) will the relation between $\langle\hat{\beta}\rangle$ and β^0 be just a scalar.

Using this linear approximation for the correlated covariate case together with the definition of the overlap parameters and $\tilde{S}^2 = \lim_{p\to\infty} \frac{1}{p}\langle \beta^0 \cdot \mathbf{A}\beta^0 \rangle_{\mathscr{D}}$, we find

$$c_0 = \lim_{p\to\infty} \frac{1}{p} \Big\langle \beta^0 \cdot \mathbf{A}\langle\beta\rangle \Big\rangle_{\mathscr{D}} = \lim_{p\to\infty} \frac{1}{p} \Big\langle \beta^0 \cdot \mathbf{A}(\kappa\beta^0 + \omega) \Big\rangle_{\mathscr{D}} = \kappa\tilde{S}^2 \tag{2.58}$$

From (2.50), $c_0 = \tilde{S}w$ hence the required slope is $\kappa = w/\tilde{S}$. Next we consider the off-diagonal elements of matrix $\mathbf{C}$

$$\begin{aligned} c &= \lim_{p\to\infty} \frac{1}{p} \Big\langle \langle\beta\rangle \cdot \mathbf{A}\langle\beta\rangle \Big\rangle_{\mathscr{D}} = \lim_{p\to\infty} \frac{1}{p} \Big\langle (\kappa\beta^0 + \omega) \cdot \mathbf{A}(\kappa\beta^0 + \omega) \Big\rangle_{\mathscr{D}} \\ &= \lim_{p\to\infty} \frac{1}{p} \Big\langle \kappa^2 \beta^0 \cdot \mathbf{A}\beta^0 + \omega \cdot \mathbf{A}\omega \Big\rangle_{\mathscr{D}} = (\kappa\tilde{S})^2 + \frac{1}{p}\mathrm{Tr}(\Omega\mathbf{A}) = (\kappa\tilde{S})^2 + \sigma^2 \end{aligned} \tag{2.59}$$

where $\langle\ldots\rangle$ represents the stochastic maximum of the penalized likelihood, $\langle\ldots\rangle_{\mathscr{D}}$ represents the average of the data and $\Omega = \sigma^2\mathbf{A}^{-1}$. The penultimate equality is calculated using

$$\begin{aligned} \langle\omega \cdot \mathbf{A}\omega\rangle_{\mathscr{D}} &= \left\langle \sum_{\mu,\nu=1}^{p} \omega_\mu A_{\mu\nu}\omega_\nu \right\rangle_{\mathscr{D}} = \sum_{\mu\nu} A_{\mu\nu}\langle\omega_\mu\omega_\nu\rangle_{\mathscr{D}} = \sum_{\mu,\nu} A_{\mu\nu}\Omega_{\mu\nu} \\ &= \sum_{\mu=1}^{p} (\mathbf{A}\Omega)_{\mu\mu} = \mathrm{Tr}(\mathbf{A}\Omega) \end{aligned} \tag{2.60}$$

The relevant order parameters from our theory are

$$c = \kappa^2\tilde{S}^2 + \sigma^2, \qquad c_0 = \kappa\tilde{S}^2 \tag{2.61}$$

Using the transformations (2.49) we obtain the following simple expressions for the two dominant characteristics κ and σ of the simulation data clouds:

$$\kappa = w/\tilde{S}, \qquad \sigma = v \tag{2.62}$$

To be clear, by assuming a model for the data cloud in (2.53), we are able to form a correspondence between our RS theory and the simulated data results. Writing the transformed order parameters $\{u, v, w\}$ in component form, we now interpret the meaning of $\{\tilde{u}, v, w\}$

$$u^2 \equiv C - c = \lim_{p\to\infty} \frac{1}{p} \sum_{\mu\nu=1}^{p} A_{\mu\nu} \langle \langle \beta_\mu \beta_\nu \rangle - \langle \beta_\mu \rangle \langle \beta_\nu \rangle \rangle_{\mathscr{D}} \tag{2.63}$$

The unscaled version, u^2 is zero in the $\gamma \to \infty$ case. However, for finite temperature, i.e. in the presence of residual noise, the value of $\tilde{u} = u\sqrt{\gamma} > 0$ is recovered from our final theory. $\tilde{u}^2$ represents the variance of the MAP-inferred $\hat{\beta}$ for a given dataset (which is subsequently averaged over all datasets).

$$v^2 \equiv c - c_0^2/\tilde{S}^2 = \lim_{p\to\infty} \frac{1}{p} \left[\sum_{\mu\nu=1}^{p} A_{\mu\nu} \langle \langle \beta_\mu \rangle \langle \beta_\nu \rangle \rangle_{\mathscr{D}} - \left(\frac{\sum_{\mu\nu=1}^{p} A_{\mu\nu} \beta_\mu^0 \langle \langle \beta_\nu \rangle \rangle_{\mathscr{D}}}{\tilde{S}} \right)^2 \right] \tag{2.64}$$

v^2 represents the variance of the MAP-inferred $\hat{\beta}$ across the whole data distribution. From (2.62), it is corresponds to the standard deviation of the data cloud in Figure 2.1. Finally the order parameter measuring the slope is

$$w = \frac{c_0}{\sqrt{C_{00}}} = \lim_{p\to\infty} \frac{1}{\sqrt{p}} \frac{\beta^0 \cdot \mathbf{A} \langle \langle \beta \rangle \rangle_{\mathscr{D}}}{\sqrt{\beta^0 \cdot \mathbf{A} \beta^0}} \tag{2.65}$$

$w/\tilde{S}$ represents the slope of the $\hat{\beta}$ versus β^0 plot. For uncorrelated covariates i.e. $\mathbf{A} = \mathbb{I}$, (2.63)-(2.65) become

$$\begin{aligned} u^2 &= \lim_{p\to\infty} \frac{1}{p} \langle \langle \beta^2 \rangle - \langle \beta \rangle^2 \rangle_{\mathscr{D}} \\ v^2 &= \lim_{p\to\infty} \frac{1}{p} \left[\langle \langle \beta \rangle^2 \rangle_{\mathscr{D}} - \left(\frac{\beta^0 \cdot \langle \langle \beta \rangle \rangle_{\mathscr{D}}}{|\beta^0|} \right)^2 \right] \\ w &= \lim_{p\to\infty} \frac{1}{\sqrt{p}} \frac{\beta^0 \cdot \langle \langle \beta \rangle \rangle_{\mathscr{D}}}{|\beta^0|} \end{aligned} \tag{2.66}$$

2.4.6 Scaling of order parameters with γ

Following [123], we scale the scalar order parameters with respect to γ

$$u = \tilde{u}/\sqrt{\gamma}, \quad v, w = \mathcal{O}(1), \quad g = \tilde{g}\gamma, \quad f = \tilde{f}\gamma^2 \tag{2.67}$$

Before taking the limit, we convert the y integral contained in Ψ_{RS}^{B} of (2.52) into a saddle point integral by defining $q \equiv y/\sqrt{\gamma}$

$$\begin{aligned}\log \int Dy\ p^{\gamma}(t|uy+wy_0+vz,r) &= \log \int \frac{\mathrm{d}y}{\sqrt{2\pi}}\ \exp\left[-\frac{1}{2}y^2+\gamma\ \log p(t|uy+wy_0+vz,r)\right] \\ &= \log \int \frac{\mathrm{d}q\sqrt{\gamma}}{\sqrt{2\pi}}\ \exp\gamma\left[-\frac{1}{2}q^2+\log p(t|\tilde{u}q+wy_0+vz,r)\right]\end{aligned} \tag{2.68}$$

Taking the limit $\gamma \to \infty$ to complete the minimization gives

$$\begin{aligned}\lim_{\gamma\to\infty}\frac{1}{\gamma}\Psi_{\mathrm{RS}}(\ldots) = \frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}+\tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle-\tilde{u}^2\right]-\tilde{g}(v^2+w^2)\right\} \\ -\int \mathrm{D}z\ \mathrm{D}y_0\int \mathrm{d}t\ p(t|\tilde{S}y_0,r_0)\ \max_q\left[\log p(t|\tilde{u}q+wy_0+vz,r)-\frac{1}{2}q^2\right]\end{aligned} \tag{2.69}$$

This expression is valid for Generalized Linear Models with a single linear predictor $\beta\cdot\mathbf{z}$. Maximization over q for the specific case of linear regression produces

$$\frac{\partial}{\partial q}[\log p(t|\tilde{u}q+wy_0+vz,r)-\frac{1}{2}q^2] = -q+\frac{\tilde{u}}{\sigma^2}[t-(\tilde{u}q+wy_0+vz+r)] \tag{2.70}$$

Resulting in the solution of the maximization $q^{\star} = \frac{\tilde{u}}{\tilde{u}^2+\sigma^2}\left[t-(wy_0+vz+r)\right]$.

In the following chapters dealing with logistic and Cox regression, we start from (2.69) and apply the relevant conditional probability with only minor changes depending on the form of the response variable t. Neglecting terms constant with respect to the order parameters,

the function to extremize is

$$\begin{aligned}\lim_{\gamma\to\infty}\frac{1}{\gamma}\Psi_{\mathrm{RS}}(\ldots) &= \frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}+\tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle-\tilde{u}^2\right]-\tilde{g}(v^2+w^2)\right\}\\ &+\frac{1}{2(\sigma^2+\tilde{u}^2)}\int \mathrm{D}z\,\mathrm{D}y_0\int \mathrm{d}t\;p(t|\tilde{S}y_0,r_0)\left[t-(wy_0+vz+r)\right]^2\\ &=\frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}+\tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle-\tilde{u}^2\right]-\tilde{g}(v^2+w^2)\right\}\\ &+\frac{1}{2(\sigma^2+\tilde{u}^2)}\left\{(\sigma^0)^2+(w-\tilde{S})^2+(r_0-r)^2+v^2\right\}\end{aligned} \tag{2.71}$$

since $p(t|\tilde{S}y_0,r_0)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}[t-(\tilde{S}y_0+r_0)]^2}$ and σ^0 is the true variance of the noise term used to generate the data.

2.4.7 Replica symmetric saddle point equations

Six order parameter equations are found by extremizing (2.71) with respect to $\{\tilde{u},v,w,r,\tilde{f},\tilde{g}\}$

$$\tilde{u}^2=\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle \tag{2.72a}$$

$$v^2=w^2\left\{\langle a\rangle\frac{\left\langle\frac{a^3}{(2\eta+\tilde{g}a)^2}\right\rangle}{\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^2}-1\right\}-\tilde{f}\left\langle\frac{a^2}{(2\eta+\tilde{g}a)^2}\right\rangle \tag{2.72b}$$

$$\zeta\tilde{f}=-\frac{1}{(\sigma^2+\tilde{u}^2)^2}\left\{(\sigma^0)^2+(w-\tilde{S})^2+(r_0-r)^2+v^2\right\} \tag{2.72c}$$

$$\zeta\tilde{g}=\frac{1}{\sigma^2+\tilde{u}^2} \tag{2.72d}$$

$$\zeta w\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}=\frac{1}{\sigma^2+\tilde{u}^2}\tilde{S} \tag{2.72e}$$

$$\frac{1}{\sigma^2+\tilde{u}^2}(r_0-r)=0 \tag{2.72f}$$

where the final equation clearly shows the inferred intercept r is equal to the true value r_0. These equations can now be solved numerically to determine values for the order parameters for a given inference problem ie $\mathbf{A}, \zeta$ and $\tilde{S}$. Instead we use (2.72a)-(2.72f) to derive a number of known statistical results for normal linear regression from our replica symmetric saddle point equations and delay numerical simulations for models where no closed form solutions are possible.

2.5 Validation of replica analysis

One may ask why this theory is necessary given that the same results are available numerically (and via classical statistical results). Firstly, by expanding our order parameter equations around $\zeta = 1$, we can find leading order behaviour exactly where numerical methods run into problems. Secondly, our theory paves the way for understanding the high-dimensional asymptotics for non-linear models of regression and classification. Lastly, analytical results can be directly compared to known statistical behaviour of the model. Now we proceed to validate our theory against the behaviour of Figure 2.1.

2.5.1 Predicted slope

Assuming uncorrelated covariates $\mathbf{A} = \mathbb{I}$ and no regularization $\eta = 0$ leads to $\tilde{u}^2 = 1/\tilde{g}$ from (2.72a), $v^2 = -\tilde{f}\tilde{u}^4$ from (2.72b), $\zeta = \frac{\tilde{u}^2}{\tilde{u}^2+\sigma^2}$ from (2.72d) and hence the slope $w/\tilde{S} = 1$ from (2.72e). The slope is independent of ζ as expected from Figure 2.1 and recovers the result that $\mathbb{E}(\hat{\beta}) = \beta^0$. We arrive at the same result by starting from the definition of the order parameter w and using (2.50), (2.58) and the maximum likelihood solution for the regression coefficients

$$\begin{aligned} w &= \lim_{p\to\infty} \frac{1}{p\tilde{S}} \left\langle \beta^0 \cdot \mathbf{A}\langle\beta\rangle \right\rangle_{\mathscr{D}} = \lim_{p\to\infty} \frac{1}{p\tilde{S}} \beta^0 \cdot \mathbf{A} \left\langle (X^T X)^{-1} X^T \mathbf{y} \right\rangle_{\mathscr{D}} \\ &= \lim_{p\to\infty} \frac{1}{p\tilde{S}} \beta^0 \cdot \mathbf{A} \left\langle (X^T X)^{-1} X^T (X\beta^0 + \varepsilon) \right\rangle_{\mathscr{D}} = \lim_{p\to\infty} \frac{1}{p\tilde{S}} \beta^0 \cdot \mathbf{A}\beta^0 = \tilde{S} \end{aligned} \tag{2.73}$$

since the noise is assumed uncorrelated with the data and $\tilde{S}^2 = \lim_{p\to\infty} \frac{1}{p} \beta^0 \cdot \mathbf{A}\beta^0$.

2.5.2 Predicted variance

Using the same assumptions as above, the slope and width of the data cloud can be expressed as

$$\tilde{u}^2 = \sigma^2 \frac{\zeta}{1-\zeta} \qquad \text{and} \qquad \mathrm{v}^2 = (\sigma^0)^2 \frac{\zeta}{1-\zeta} \tag{2.74}$$

Limit $\zeta \to 0$. To equate the expressions when $N \gg p$, we use an alternative expression for variance of the inferred regression coefficients [110]

$$\mathrm{Var}(\hat{\beta}_\mu) = \left[\sigma^2 (\mathbf{Z}^T\mathbf{Z})^{-1}\right]_{\mu\mu} = \frac{\sigma^2}{(N-1)\mathrm{Var}(z_\mu)} \mathrm{VIF}_\mu \tag{2.75}$$

where $\mathrm{Var}(z_\mu)$ is the empirical variance of the μ^{th} covariate and VIF_μ is the variance inflation factor for covariate μ defined in (2.11). Hence in the limit $N \to \infty$ for a given model (p fixed), $\mathrm{Var}(\hat{\beta}_\mu) \to 0$ in agreement with (2.74) as $\zeta \to 0$.

Limit $\zeta \to 1$. Both v^2 and $\mathrm{Var}(\hat{\beta}_\mu) = \sigma^2 \left[(\mathbf{Z}^T\mathbf{Z})^{-1}\right]_{\mu\mu}$ are proportional to σ^2 and are undefined when $p \geq N$ i.e. $\zeta = 1$.

Next we take the classical expression, average over the data and take the limit $p, N \to \infty$ to allow comparison with (2.74). The average variance over all regression parameters is

$$\frac{1}{p}\sum_{\mu=1}^{p} \mathrm{Var}(\hat{\beta}_\mu) = \sigma^2 \frac{1}{p}\sum_{\mu=1}^{p} \left[(\mathbf{Z}^T\mathbf{Z})^{-1}\right]_{\mu\mu} = \sigma^2 \frac{1}{p}\mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{-1} \to \sigma^2 \int \mathrm{d}\lambda\ \rho_{\mathrm{MP}}(\lambda)\ \frac{1}{\lambda} \tag{2.76}$$

where $\{\lambda_\mu\}_{\mu=1}^p$ are the eigenvalues of matrix $\mathbf{Z}^T\mathbf{Z}$ and the required limiting eigenvalue spectrum is given by the Marčenko-Pastur equation $\rho_{\mathrm{MP}}(\lambda)$. We differ from (1.15) by a factor of ζ since we are no longer scaling the original covariance matrix by N^{-1}. We are left to calculate

$$\int \mathrm{d}\lambda\ \rho_{\mathrm{MP}}(\lambda)\ \frac{1}{\lambda} = \int \mathrm{d}\lambda\ \frac{\sqrt{(\lambda-\lambda_{\min})(\lambda_{\max}-\lambda)}}{2\pi\lambda^2} \tag{2.77}$$

where $\lambda = [\lambda_{\min}, \lambda_{\max}]$ and $\lambda_{\min} = \left(1-\sqrt{\zeta}\right)^2$ and $\lambda_{\max} = \left(1+\sqrt{\zeta}\right)^2$ and $\rho_{\mathrm{MP}}(\lambda)$ is defined for $\zeta \leq 1$. Clearly the eigenvalues lie on the positive real line as expected for a covariance matrix. We find that in the limit $p \to \infty$ with $p/N \sim \mathscr{O}(1)$, the predicted variance from the ML order parameters equations (2.74) is equivalent to the classical result:

$$\left\langle \mathrm{Var}(\hat{\beta}) \right\rangle_{\mathscr{D}} = \sigma^2 \left\langle \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})^{-1} \right\rangle_{\mathscr{D}} \to \sigma^2 \frac{\zeta}{1-\zeta} \tag{2.78}$$

See Appendix B.6.2 for details.

Phase transition. The order parameters $\tilde{u}$ and v diverge as $\zeta \to 1$. Hence our theory in the maximum likelihood case agrees with the classical result on the location of the phase transition, $\zeta = 1$. In addition, setting the regularizer $\eta = 0$ (maximum likelihood) eliminates the dependence on the eigenvalues of $\mathbf{A}$ in (2.72a), (2.72b). This is shown analytically in appendix B.1.

2.6 Summary

In this chapter, we developed a theory for statistical inference applied to the normal linear regression model. Many calculations could be performed analytically allowing us to focus on the methodological detail. Reassuringly our results agreed with classical statistics in the $p \ll N$ regime but were also valid when $p \sim N$. In this latter regime, traditional results are generally lacking. This statistical physics framework generalized previous work from ML [26] to MAP estimation and additionally allowed for the use of correlated covariates. Also expressions for the macroscopic parameters of the inference problem could be calculated directly without the need for the replica method.

In the following chapters, we study binary classification and time-to-event models with the statistical physics formalism developed. We will find that closed form solutions for the model parameters are no longer available and that the direct calculation approach of Section 2.3 does not simplify the problem. Further, we find surprising reproducible effects on the macroscopic parameters as p/N increases.

Chapter 3

Regularized Logistic Regression

3.1 Introduction

In the previous chapter, we considered the continuous response variable of normal linear regression. Now we apply our replica formalism to the binary response of logistic regression. The slope and width of the resulting regression outcomes are again investigated. In fact, we will see that our theory can be applied to a range of conditional probabilities with a single linear predictor $\beta \cdot \mathbf{z}$ and that the non-linear link function results in interesting behaviour even in the maximum likelihood case. Differences arise in the minimization which no longer results in analytically tractable expressions and leads to a more computationally expensive calculation of the order parameters. Simulations produce a result not seen in linear regression inspiring the development of our current method.

In contrast to the linear model (Figure 2.1), we find the slope of the data cloud formed during inference using logistic regression depends on ζ. In this chapter, we apply our theory to this non-linear GLM in order to predict the systematic biases suggested from the simulations (Figure 3.1) in the regime $p \sim \mathcal{O}(N)$. We note there is surprisingly little theoretical work available on this effect. By extending our theory to logistic regression and varying values of ζ and r^0, we find explanations for the unwelcome behaviour of this binary classifier under class imbalance. Both of these topics are particularly relevant in the modern-era of data collection.

The model. Logistic regression, with its non-linear link function $g(x) = \log\ [x/(1-x)]$, is used as a model of binary classification. The conditional probability with response variables $t \in \{-1, +1\}$, which differs from the real-valued response variables in linear regression,

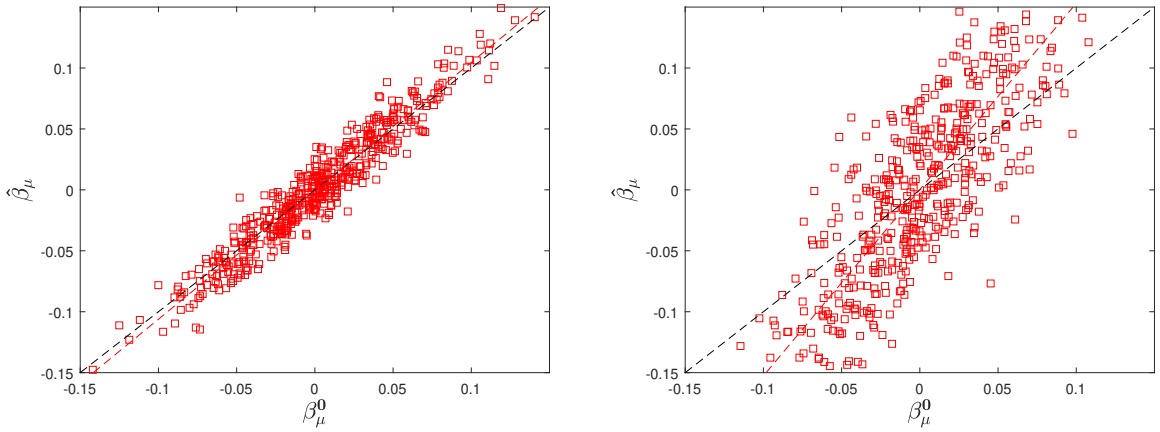

Fig. 3.1 Comparison of true β^0 and inferred $\hat{\beta}$ regression coefficients for the Logistic Regression model. Synthetic data was generated using Gaussian covariates with $p = 500$ and two values of $N = 10,000$ and $N = 2,500$ resulting $\zeta = 0.05$ (left) and $\zeta = 0.20$ (right). The red dashed line is the best fit to the data cloud. The regression coefficients are estimated via maximum likelihood inference (using the R package *glmnet* [56]). The black dashed line $\hat{\beta} = \beta^0$ represents perfect inference. Both the slope and variance of the data cloud increases with ζ.

becomes

$$p(t|\mathbf{z},r,\beta) = \frac{e^{t(r+\beta\cdot\mathbf{z})}}{2\cosh(r+\beta\cdot\mathbf{z})} = \frac{1}{2}\left[1+\tanh t(r+\beta\cdot\mathbf{z})\right] \tag{3.1a}$$

$$\log p(t|\mathbf{z},r,\beta) = t(r+\beta\cdot\mathbf{z}) - \log 2\cosh(r+\beta\cdot\mathbf{z}) \tag{3.1b}$$

The model parameters are $\vartheta = \{r,\beta\}$ where the intercept term $r \in \mathbb{R}$ and the regression coefficients $\beta = \{\beta_1,\ldots,\beta_p\} \in \mathbb{R}^p$. As before $\{r^0,\beta^0\}$ are the true model parameters used to generate the data.

Although we will not use it in our analysis, an alternative form of the conditional probability distribution is shown for completeness. It is familiar to the statistics community and can be related to (3.1a) using the standard hyperbolic identities

$$p(t|\mathbf{z},r,\beta) = \left(\frac{1}{1+e^{-(r+\beta\cdot\mathbf{z})}}\right)^t \left(\frac{1}{1+e^{(r+\beta\cdot\mathbf{z})}}\right)^{1-t} \tag{3.2}$$

The response variable $t \in \{0,1\}$ can be transformed to the $\{-1,+1\}$ formulation typical of statistic mechanics problems.

As with the linear case, we intentionally separate the intercept term r from the other parameters. This will become useful when considering class imbalance which is controlled by the intercept term (also known as the bias or threshold).

Two branches of literature have given rise to distinct methodologies for analyzing the logistic regression problem: binary classification and the perceptron learning problem.

Binary classification. Statistical methods. Logistic regression was first used in the nineteenth century for applications in population growth and chemical reactions [144] typically for $N \gg p$. More recently, biomedical applications have ranged from estimation of propensity scores[1] [112] to the classification of gene expression data [149]. Increased data-gathering and computational power generated a need to accommodate problems where $p \sim \mathscr{O}(N)$. Attempts to control the magnitude or sparsity of inferred parameters led to the development of the so-called regularized logistic regression model. See [55] for an overview of these methods.

Perceptron. Statistical physics methods. The perceptron (see Figure 3.2) uses an activation function, such as $\sigma(x) = [1+\exp(-x)]^{-1}$, to convert p inputs[2] into a single response variable. It was introduced by Rosenblatt [113] who provided a convergence proof of an associated learning algorithm for linearly separable data [114]. Following the work of [57] which calculated the maximum number of patterns storable, the analysis of the perceptron as a learning tool enjoyed a resurgence of interest. Researchers used replica theory [115], dynamical replica theory [28] and generating functional analysis [69] to tackle the challenging $p \sim \mathscr{O}(N)$ regime for both equilibrium and dynamical settings. Results were obtained for combinations of perceptrons e.g. committee machines [29, 119].

Recently tools of approximate message passing [97] have been used to study logistic regression in the high-dimensional regime for uncorrelated covariates. The ML case was covered in [133] and the regularized logistic regression model in [117] (both references assuming uncorrelated covariate inputs). In the unregularized model, correlated covariates can be harmlessly transformed to uncorrelated ones but introducing a regularizer means this is no longer holds. Implications for likelihood ratio tests in the $p \sim N$ regime are explored in [134].

The majority of the literature considers the student-teacher scenario where the teacher weights (or ϑ^0 in our notation) are fixed and used to generate the data. This is given to

[1] The propensity score is the conditional probability of a patient receiving treatment given a set of covariates. It is used to mitigate confounding when considering the effect of medical treatment on patient outcomes.

[2] In the perceptron literature the dimension of input vectors is typically labelled as n and the number of patterns as p. In this paper, we follow statistics notation of N samples and p dimensions. This is consistent with the rest of this thesis.

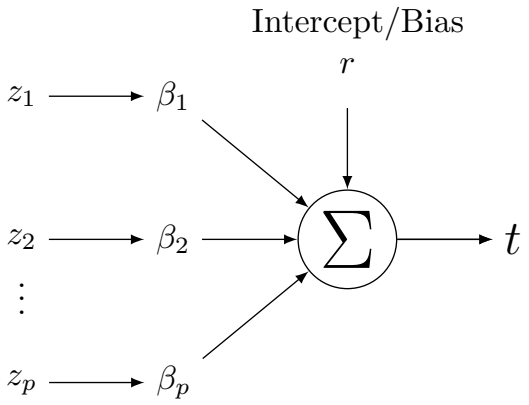

Fig. 3.2 A perceptron with p inputs $\{z_1,\ldots,z_p\}$, weight vector β, threshold term r and response variable t.

the student along with the correct model (but importantly not the weights) and it is left to the student to estimate the weights with the measure of success typically being the overlap between the student and teacher weight vectors. Both dynamic and equilibrium results have been considered. This scenario is also known as the *Bayes optimal* and is a realizable problem. Using the notation of (1.10), $\exists\beta \in \mathbb{R}^p$ such that $\varepsilon(\beta;\mathbf{z}_i,t_i)=0, \quad \forall i=\{1,\ldots,N\}$.

What have we added to this vast literature on the perceptron? A specific case of correlated variables controlled by the overall "magnetization" variable was considered in [57]. As we have seen from the previous chapter, our replica formalism is valid for a general covariance matrix with mild conditions on the eigenvalue spectrum. In addition, we apply our analysis to the regularized logistic regression model. Lastly, by separating out the intercept term, we investigate the phenomenon of misclassification under class imbalanced data in high dimensions.

3.1.1 Intuition

The information-theoretic overfitting measure introduced in (2.23) can be rewritten, in the absence of regularization, using the logistic regression form (3.1a) to

$$\mathrm{E}(\beta^0,\mathscr{D}) = \frac{1}{N}\sum_{i=1}^{N}\left\{ t_i\left[(r^0+\beta^0\cdot\mathbf{z}_i)-(r+\beta\cdot\mathbf{z}_i)\right]+\log\frac{\cosh(r+\beta\cdot\mathbf{z}_i)}{\cosh(r^0+\beta^0\cdot\mathbf{z}_i)}\right\} \tag{3.3}$$

or alternatively using the formulation of (3.2)

$$\mathrm{E}(\beta^0, \mathscr{D}) = \frac{1}{N}\sum_{i=1}^{N}\left\{ t_i \log\left(\frac{1+\mathrm{e}^{-\mathrm{r}-\beta\cdot\mathbf{z}_\mathrm{i}}}{1+\mathrm{e}^{-\mathrm{r}^0-\beta^0\cdot\mathbf{z}_\mathrm{i}}}\right) + (1-t_i)\log\left(\frac{1+\mathrm{e}^{\mathrm{r}+\beta\cdot\mathbf{z}_\mathrm{i}}}{1+\mathrm{e}^{\mathrm{r}^0+\beta^0\cdot\mathbf{z}_\mathrm{i}}}\right)\right\} \tag{3.4}$$

The previous chapter minimizes E analytically for normal linear regression. To gain intuition for the overfitting measure using the current model (3.4), we carry out the minimization numerically. Pseudo-code for the simulation protocol is given in Algorithm 1 for reproducibility. The Nelder-Mead algorithm [102] was implemented using the *optimr* package in R [101].

Algorithm 1 Protocol for estimating the overfitting measure

1: **procedure** OVERFITTING PROTOCOL
2: Generate $\beta^0, \mathbf{z}_i \sim \mathrm{N}_p(0, \mathbb{I})$ and $\mathbf{y} \in \mathbb{R}$ from (3.2)
3: **while** Nelder-Mead iteration **do**
4: Update maximum likelihood estimate $\hat{\beta}$
5: calculate E from (3.4)
6: **end while**
7: Plot E versus iteration steps.
8: **end procedure**

Other search methods for finding the inferred regression parameters, such as stochastic gradient descent or Iteratively Reweighted Least Squares, have different advantages depending on the data for a convex optimization problem. Since our overfitting measure is a convex function of β (see Appendix B.5), the same solution will be reached albeit at different rates. Simulation results are shown in Table 3.1 and Figure 3.3.

N	p	ζ	$\mathrm{E}_{\mathrm{final}}$
10	25	2.5	-0.33
100	25	0.25	-0.16
1000	25	0.025	0.02

Table 3.1 Overfitting measure for various values of N and $p = 25$

As expected when ζ is close to zero (0.025), the overfitting measure E converges towards zero during the minimization. As the number of samples in the data set is reduced ($\zeta = 0.25, 2.5$), E converges to increasingly negative values. Since there is no model mismatch (the data were generated from a logistic model), the negative values of E indicate overfitting.

It is tempting to imagine that the inferred regression coefficients $\hat{\beta}$ corresponding to E crossing zero in Figure 3.3 should result in perfect inference. Informally, values of $\beta \in \mathbb{R}^p$ which satisfy $\mathrm{E}(\beta^0, \mathscr{D}) = 0$ in (3.4) lie on a p-dimensional sphere around the empirical data

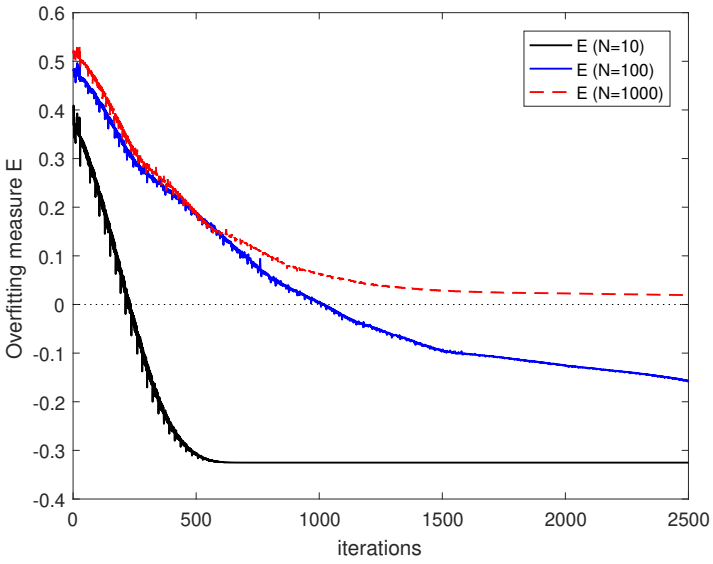

Fig. 3.3 Synthetic data with dimension $p = 25$ and $N = \{10, 100, 1000\}$ are generated using the logistic regression model according to Algorithm 1. The ML estimate of model parameters is found numerically using the Nelder-Mead algorithm and the overfitting measure E plotted after each iteration. The starting value model parameters β in the minimization search is the zero vector, giving an initial positive value of E (implying an underfitted model).

distribution. The point where the minimization algorithm meets this sphere is a random variable dependent on the initialization vector (and the algorithm itself) and is therefore not guaranteed to be the true value β^0. As a numerical experiment, we record the value of the normalized scalar product $(\beta \cdot \hat{\beta})/|\beta||\hat{\beta}|$ when the overfitting measure crosses zero and its final value (for two values of ζ). The majority of markers are above the diagonal in Figure 3.4 representing simulations where the ML estimator is a better estimate than the intermediate β value when $\mathrm{E} = 0$. Put another way, since (2.23) is defined as a minimization, intermediate stages of the iteration are not optimal.

3.2 Replica theory

3.2.1 Model definition

Having gained some intuition for the overfitting measure, we proceed to calculate relevant properties of the inference problem using the replica formalism developed previously. We find that calculations are almost identical to the linear regression case up to (2.69) where the replica symmetric ansatz has been assumed and the N and n limits have been taken. Only model-specific complications are left to deal with. Differences in the response variable t are

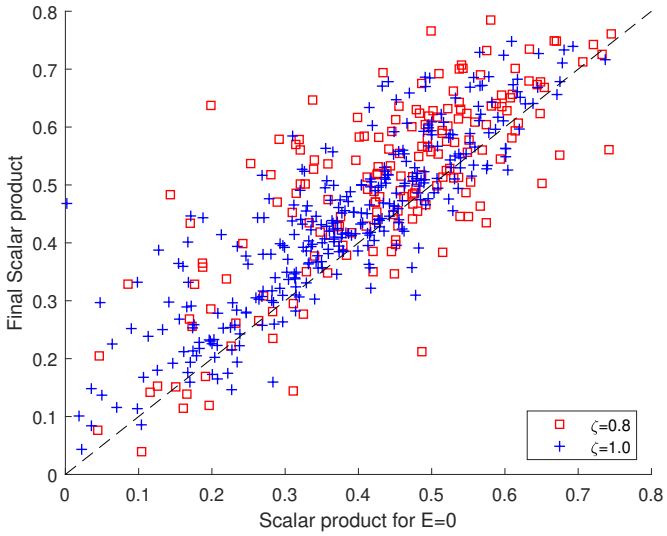

Fig. 3.4 Maximum likelihood estimation for $\zeta = \{0.8, 1.0\}$ with $N = 25$. Markers above the diagonal represent simulations where the final scalar product $\hat{\beta}_{\mathrm{ML}}$ is closer than $\hat{\beta}_{\mathrm{E}=0}$ to the true β^0.

highlighted by repeating the initial steps from Section 2.4. The free energy expression is

$$\mathrm{E}_\gamma(\vartheta^0) = -\frac{1}{N}\frac{\partial}{\partial\gamma}\left\langle \log \int \mathrm{d}\vartheta \prod_{i=1}^{N} \left[\frac{p(t_i|\mathbf{z}_i,\vartheta)\; p(\vartheta)}{p(t_i|\mathbf{z}_i,\vartheta^0)\; p(\vartheta^0)}\right]^\gamma \right\rangle_{\mathscr{D}} \tag{3.5}$$

Inserting the conditional probability (3.1a) into (3.5), using the replica identity $\langle \log Z\rangle = \lim_{n\to 0}\; n^{-1}\log\langle Z^n\rangle$ and noting the integral over the response variable $t \in \mathbb{R}^p$ in (2.26) has become a sum over the discrete response variable $t = \{-1, +1\}$.

$$\begin{aligned}\mathrm{E}_\gamma(\beta^0, r^0) = -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log \int \mathrm{d}r^1 \ldots \mathrm{d}r^n \int \mathrm{d}\beta^1 \ldots \mathrm{d}\beta^n \left\{\prod_{\alpha=1}^{n}\left[\frac{p(\beta^\alpha)}{p(\beta^0)}\right]^\gamma\right\} \\ \times\left\{\int \sum_{t=\pm 1} \mathrm{d}\mathbf{z}\; p(\mathbf{z})\, p(t|\mathbf{z}, r^0, \beta^0) \prod_{\alpha=1}^{n}\left[\frac{p(t|\mathbf{z}, r^\alpha, \beta^\alpha))}{p(t|\mathbf{z}, r^0, \beta^0)}\right]^\gamma\right\}^N\end{aligned} \tag{3.6}$$

where $\{r^0, \beta^0\}$ represent the true regression and intercept parameters and $\{r^\alpha, \beta^\alpha\}$ the parameters inferred via MAP estimation. The index over independent replicas of the system is $\alpha \in \{1, \ldots, n\}$.

To proceed, we follow the same steps as the linear regression case by introducing the delta function, conversion to a saddle point problem and assuming the replica symmetric ansatz. These manipulations result in (2.69) for the linear regression case and after appropriate scaling (3.8) in the current logistic regression case.

3.2.2 Taking the $\gamma \to \infty$ limit

Recalling the scaling of the order parameters with respect to γ in the linear case

$$u = \tilde{u}/\sqrt{\gamma}, \quad v, w = \mathcal{O}(1), \quad g = \tilde{g}\gamma, \quad f = \tilde{f}\gamma^2 \tag{3.7}$$

To complete the minimization, we take the limit $\gamma \to \infty$ and note the t integral over the real line becomes a sum

$$\begin{aligned}\lim_{\gamma\to\infty} \frac{1}{\gamma}\Psi_{\mathrm{RS}}(\ldots) = \frac{1}{2}\zeta\left\{ w^2\langle a\rangle \Big\langle \frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-1} + \tilde{f}\Big[\Big\langle \frac{a}{2\eta+\tilde{g}a}\Big\rangle - \tilde{u}^2\Big] - \tilde{g}(v^2+w^2)\right\} \\ - \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\, \max_{q\in\mathbb{R}}\Big[\log p(t|\tilde{u}q + wy_0 + vz, r) - \frac{1}{2}q^2\Big]\end{aligned} \tag{3.8}$$

where we have used the same transformation $q \equiv y/\sqrt{\gamma}$ as Section 2.4.6 to facilitate the $\gamma \to \infty$ limit. Inserting the conditional probability for logistic regression (3.1b) and defining

$$\begin{aligned}\Phi[q(\tilde{u}, v, w, r)] \equiv\ & \max_{q\in\mathbb{R}}\Big[-\frac{1}{2}q^2 + \log p(t|\tilde{u}q + wy_0 + vz + r)\Big] \\ =\ & \max_{q\in\mathbb{R}}\Big[-\frac{1}{2}q^2 + t(\tilde{u}q + wy_0 + vz + r) - \log\ 2\cosh(\tilde{u}q + wy_0 + vz + r)\Big] \\ =\ & \max_{q\in\mathbb{R}}\ \phi(q)\end{aligned} \tag{3.9}$$

Maximizing $\phi(q)$ with respect to q results in the following transcendental equation

$$\frac{\partial\phi(q)}{\partial q} = 0: \qquad q = \tilde{u}t - \tilde{u}\tanh(\tilde{u}q + wy_0 + vz + r) \tag{3.10}$$

The solution of (3.10) is a function of the order parameters i.e. $q = q(\tilde{u}, v, w, r)$. What remains in our RS analysis is to determine the order parameter equations by extremization of

$$\begin{aligned}\lim_{\gamma\to\infty} \frac{1}{\gamma}\Psi_{\mathrm{RS}}(\ldots) = \frac{1}{2}\zeta\left\{ w^2\langle a\rangle \Big\langle \frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-1} + \tilde{f}\Big[\Big\langle \frac{a}{2\eta+\tilde{g}a}\Big\rangle - \tilde{u}^2\Big] - \tilde{g}(v^2+w^2)\right\} \\ - \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\, \Phi[q(\tilde{u}, v, w, r)]\end{aligned} \tag{3.11}$$

We note the split into the model independent first line of (3.11) and the model specific term in the second line. This highlights the general nature of our formulation (see (2.52) for the linear regression equivalent).

3.2.3 Differentiation with respect to order parameters

Differentiation with respect to the conjugate order parameters $\tilde{f}$ and $\tilde{g}$ result in identical equations to the linear regression case i.e. (2.72a) and (2.72b). Since q is a function of the order parameters, its derivative is first found with respect to the relevant order parameter from the implicit equation (3.10) and then the derivative of Φ taken. We use (3.10) repeatedly in the following derivations in particular $q^\star/\tilde{u} = t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)$.

Differentiation with respect to $\tilde{u}$
From (3.10),

$$\frac{\partial q^\star}{\partial \tilde{u}} = t - \tilde{u}(1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r))\big(\tilde{u}\frac{\partial q^\star}{\partial \tilde{u}} + q^\star\big) - \tanh(\tilde{u}q^\star + wy_0 + vz + r) \tag{3.12}$$

$$\frac{\partial q^\star}{\partial \tilde{u}} = \frac{q^\star}{\tilde{u}} - \tilde{u}\left(\tilde{u}\frac{\partial q^\star}{\partial \tilde{u}} + q^\star\right)[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)] \tag{3.13}$$

Re-arranging to find an explicit expression

$$\frac{\partial q^\star}{\partial \tilde{u}} = \frac{(q^\star/\tilde{u}) - \tilde{u}q^\star[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)]}{1 + \tilde{u}^2[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)]} \tag{3.14}$$

Now evaluate

$$\begin{aligned}\frac{\partial}{\partial \tilde{u}}\Phi[q(\tilde{u},v,w,r)] &= -q^\star\frac{\partial q^\star}{\partial \tilde{u}} + t\left(\tilde{u}\frac{\partial q^\star}{\partial \tilde{u}} + q^\star\right) - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\left(\tilde{u}\frac{\partial q^\star}{\partial \tilde{u}} + q^\star\right) \\ &= -q^\star\frac{\partial q^\star}{\partial \tilde{u}} + \left(\tilde{u}\frac{\partial q^\star}{\partial \tilde{u}} + q^\star\right)\frac{q^\star}{\tilde{u}} = \frac{(q^\star)^2}{\tilde{u}}\end{aligned} \tag{3.15}$$

The order parameter equation becomes

$$\begin{aligned}\zeta\tilde{f}\tilde{u} &= -\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\frac{\partial}{\partial \tilde{u}}\Phi[q(\tilde{u},v,w,r)] \\ &= -\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\frac{(q^\star)^2}{\tilde{u}}\end{aligned} \tag{3.16}$$

Using $\frac{q^\star}{\tilde{u}} = t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)$ and $t^2 = 1$

$$\begin{aligned}
\zeta \tilde{f} &= -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left(\frac{q^\star}{\tilde{u}}\right)^2 \\
&= -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right]^2 \\
&= -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[1 - \tanh t(\tilde{u}q^\star + wy_0 + vz + r)\right]^2
\end{aligned} \tag{3.17}$$

The following two calculations involves the integration by parts and hence the second derivative with respect to the order parameters.

Differentiation with respect to v
Similarly to the $\tilde{u}$ derivation, we find

$$\frac{\partial q^\star}{\partial v} = \frac{-\tilde{u}z\left[1 - \tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]}{1 + \tilde{u}^2\left[1 - \tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]} \tag{3.18}$$

Now evaluate

$$\begin{aligned}
\frac{\partial}{\partial v}\Phi[q(\tilde{u}, v, w, r)] &= -q^\star \frac{\partial q^\star}{\partial v} + \left(\tilde{u}\frac{\partial q^\star}{\partial v} + z\right)\left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] \\
\frac{\partial}{\partial v}\Phi[q(\tilde{u}, v, w, r)] &= -q^\star \frac{\partial q^\star}{\partial v} + \left(\tilde{u}\frac{\partial q^\star}{\partial v} + z\right)\frac{q^\star}{\tilde{u}} \\
&= z\left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right]
\end{aligned} \tag{3.19}$$

The relevant order parameter equation becomes

$$\begin{aligned}
\zeta \tilde{g}v &= -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \frac{\partial}{\partial v}\Phi[q(\tilde{u}, v, w, r)] \\
&= -\int \mathrm{D}y_0 \mathrm{D}z\, z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] \\
&= -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \frac{\partial}{\partial z}\left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] \\
&= \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[1 - \tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]\left(\tilde{u}\frac{\partial q^\star}{\partial z} + v\right)
\end{aligned} \tag{3.20}$$

Finally

$$\zeta\tilde{g} = \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \frac{1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)}{1+\tilde{u}^2[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)]} \tag{3.21}$$

Differentiation with respect to w

We find

$$\frac{\partial q^\star}{\partial w} = \frac{-\tilde{u}y_0\left[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]}{1+\tilde{u}^2\left[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]} \tag{3.22}$$

Now evaluate

$$\begin{aligned}\frac{\partial}{\partial w}\Phi[q(\tilde{u},v,w,r)] &= -q^\star\frac{\partial q^\star}{\partial w} + \left(\tilde{u}\frac{\partial q^\star}{\partial w} + y_0\right)\Big[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\Big] \\ &= y_0\ \left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right]\end{aligned} \tag{3.23}$$

Here the integration by parts contains the product of two functions of y_0. Re-writing $p(t|\tilde{S}y_0, r_0) = \frac{1}{2}[1+\tanh t(\tilde{S}y_0 + r_0)]$

$$\begin{aligned}&\zeta w\left\{\tilde{g} - \langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}\right\} \\ &= -\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\frac{\partial}{\partial w}\Phi[q(\tilde{u},v,w,r)] \\ &= -\int \mathrm{D}y_0\mathrm{D}z\, y_0 \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] \\ &= -\frac{1}{2}\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1}\frac{\partial}{\partial y_0}[1+\tanh t(\tilde{S}y_0 + r_0)][t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)] \\ &= \frac{1}{2}\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1}\Big\{[1+\tanh t(\tilde{S}y_0 + r_0)]\frac{1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)}{1+\tilde{u}^2[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)]}w \\ &\qquad - t\tilde{S}[1-\tanh^2 t(\tilde{S}y_0 + r_0)][t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)]\Big\}\end{aligned} \tag{3.24}$$

Combining with (3.21) leads to cancelling the $\zeta w\tilde{g}$ terms:

$$\begin{aligned}\zeta w\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1} &= \frac{1}{2}\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} t\tilde{S}[1-\tanh^2 t(\tilde{S}y_0 + r_0)][t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)] \\ &= \frac{1}{2}\tilde{S}\int \mathrm{D}y_0\mathrm{D}z\ [1-\tanh^2(\tilde{S}y_0 + r_0)] \sum_{t=\pm 1}[1-\tanh t(\tilde{u}q^\star + wy_0 + vz + r)]\end{aligned} \tag{3.25}$$

Differentiation with respect to r

Finally

$$\frac{\partial q^\star}{\partial r} = \frac{-\tilde{u}\left[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]}{1+\tilde{u}^2\left[1-\tanh^2(\tilde{u}q^\star + wy_0 + vz + r)\right]} \tag{3.26}$$

Now evaluate

$$\begin{aligned}
\frac{\partial}{\partial r}\Phi[q(\tilde{u},v,w,r)] &= -q^\star\frac{\partial q^\star}{\partial r} + \left(\tilde{u}\frac{\partial q^\star}{\partial r}+1\right)\left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] \\
&= -q^\star\frac{\partial q^\star}{\partial r} + \left(\tilde{u}\frac{\partial q^\star}{\partial r}+1\right)\frac{q^\star}{\tilde{u}} \\
&= t - \tanh(\tilde{u}q^\star + wy_0 + vz + r) \\
&= \frac{q^\star}{\tilde{u}}
\end{aligned} \tag{3.27}$$

The order parameter equation becomes

$$\begin{aligned}
0 &= -\int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\frac{\partial}{\partial r}\Phi[q(\tilde{u},v,w,r)] \\
&= \int \mathrm{D}y_0\mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\; q^\star
\end{aligned} \tag{3.28}$$

We note from (3.10) that $q(\tilde{u},v,w,r)$ is a function of the Gaussian variables $\{y_0,z\}$. Assuming $r^0 = 0$ implies $\int \mathrm{D}y_0\,[1+\tanh(\tilde{S}y_0+r_0)] = \int \mathrm{D}y_0\,[1-\tanh(\tilde{S}y_0+r_0)]$ hence $r^0 = 0 \;\Rightarrow\; r = 0$.

3.2.4 Replica symmetric saddle point equations

The six order parameter equations, found via $\nabla\Psi_{RS} = \mathbf{0}$, define the relationship between $\{\tilde{u}, v, w, r, \tilde{f}, \tilde{g}\}$

$$\tilde{u}^2 = \left\langle \frac{a}{2\eta + \tilde{g}a} \right\rangle \tag{3.29a}$$

$$v^2 = w^2 \left\{ \langle a \rangle \frac{\left\langle \frac{a^3}{(2\eta+\tilde{g}a)^2} \right\rangle}{\left\langle \frac{a^2}{2\eta+\tilde{g}a} \right\rangle^2} - 1 \right\} - \tilde{f} \left\langle \frac{a^2}{(2\eta + \tilde{g}a)^2} \right\rangle \tag{3.29b}$$

$$\zeta \tilde{f} = -\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[1 - \tanh\, t(\tilde{u}q^\star + wy_0 + vz + r)\right]^2 \tag{3.29c}$$

$$\zeta \tilde{g} = \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} \frac{p(t|\tilde{S}y_0, r_0)}{\tilde{u}^2 + \cosh^2(\tilde{u}q^\star + wy_0 + vz + r)} \tag{3.29d}$$

$$\zeta w \langle a \rangle \left\langle \frac{a^2}{2\eta + \tilde{g}a} \right\rangle^{-1} = \frac{1}{2}\tilde{S} \int \mathrm{D}y_0 \mathrm{D}z \, [1 - \tanh^2(\tilde{S}y_0 + r_0)] \sum_{t=\pm 1} [1 - \tanh t(\tilde{u}q^\star + wy_0 + vz + r)] \tag{3.29e}$$

$$\int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \left[t - \tanh(\tilde{u}q^\star + wy_0 + vz + r)\right] = 0 \tag{3.29f}$$

Maximum Likelihood equations. By assuming the simplest case of uncorrelated covariates $\mathbf{A} = \mathbb{I}$ and no regularization $\eta = 0$ leads to $\tilde{u}^2 = 1/\tilde{g},\ v^2 = -\tilde{f}\tilde{u}^4$. Further, if we assume the

true intercept $r^0 = 0$, we can set $r = 0$. The three remaining equations to be solved become

$$\zeta v^2 = \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0)\ (\tilde{u}q^\star)^2 \tag{3.30a}$$

$$\zeta = \int \mathrm{D}y_0 \mathrm{D}z \sum_{t=\pm 1} p(t|\tilde{S}y_0, r_0) \frac{\tilde{u}^2}{\tilde{u}^2 + \cosh^2(\tilde{u}q^\star + wy_0 + vz + r)} \tag{3.30b}$$

$$\zeta w = \frac{1}{2}\tilde{S} \int \mathrm{D}y_0 \mathrm{D}z\ [1 - \tanh^2(\tilde{S}y_0 + r_0)] \sum_{t=\pm 1} t\tilde{u}q^\star \tag{3.30c}$$

It is not clear how to proceed with the summations over the response variables since $q^\star$ is an implicit function of t. By making the following transformation, the t dependency becomes explicit allowing simplifications in preparation for numerical simulations.

$$x \equiv t(\tilde{u}q^\star + wy_0 + vz) \qquad \Rightarrow \qquad x - t(wy_0 + vz) = t\tilde{u}q^\star \tag{3.31}$$

Transforming the Gaussian variables $\{y_0, z\} \to \{-y_0, -z\}$ in one of the summation terms and noting $x = x(\tilde{u}, v, w, y_0, z)$, we find

$$\zeta v^2 = \int \mathrm{D}y_0 \mathrm{D}z\ [1 + \tanh(\tilde{S}y_0)]\ [x - (wy_0 + vz))]^2 \tag{3.32a}$$

$$\zeta = \int \mathrm{D}y_0 \mathrm{D}z\ [1 + \tanh(\tilde{S}y_0)]\ \frac{\tilde{u}^2}{\tilde{u}^2 + \cosh^2(x)} \tag{3.32b}$$

$$\zeta w = \tilde{S} \int \mathrm{D}y_0 \mathrm{D}z\ [1 - \tanh^2(\tilde{S}y_0 + r_0)]\ x \tag{3.32c}$$

Reassuringly, solutions of (3.32a)-(3.32c) agree with numerical simulations. These are shown in the following section along with results for the general order parameter equations (3.29a)-(3.29f) (see Figures 3.5-3.7).

3.3 Numerical results

For the linear regression model, analytical expressions derived from the relevant saddle point equations were validated against known statistical results (2.7), (2.9). In contrast, attempts to find closed form solutions for $\hat{\beta}$ in the logistic regression model lead to a system of N

non-linear equations.

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{N} \frac{e^{t_i(r+\beta\cdot\mathbf{z}_i)}}{2\cosh(r+\beta\cdot\mathbf{z}_i)} \\
\ell(\beta) &= \sum_{i=1}^{N} \left[t_i(r+\beta\cdot\mathbf{z}_i) - \log 2\cosh(r+\beta\cdot\mathbf{z}_i)\right] \\
\frac{\partial}{\partial\beta_j}\ell(\beta) = 0: &\qquad \sum_{i=1}^{N} z_{ij}\,\left[t_i - \tanh(r+\hat{\beta}\cdot\mathbf{z}_i)\right] = 0
\end{aligned}
\tag{3.33}
$$

where z_{ij} is the j^{th} component of vector $\mathbf{z}_i$. These transcendental equations do not have a closed form solution so numerical methods are typically used to estimate the model parameters. In this spirit, we first compare our theory to synthetic data in ML regression. This has the benefit of allowing comparisons to [133] where equivalent terms are derived using an approximate message passing algorithm. Next we repeat these comparisons for the regularized case with uncorrelated data.

3.3.1 ML case

Equations (3.32a)-(3.32c) relate $\{\tilde{u}, v, w\}$ in the simplest case of $\mathbf{A} = \mathbb{I},\ \eta = 0$ and $r^0 = 0$ and contain information about the location of the phase transition in the statistical physics problem. This will help us identify the corresponding region of the original inference problem which are difficult. Recall that in the normal linear model, the location of the phase transition is always at $\zeta = 1$ irrespective of values of $\tilde{S}$. For the logistic regression model, the ground-breaking work of [31, 57] may have led some to the assumption that the phase transition is always at $\zeta = 0.5$. Numerical solutions to (3.32a)-(3.32c), along with simulated results, are shown in Figure 3.5 for a range of $\tilde{S}$ values. Deriving analytical expressions for the location of the phase transition in terms of ζ and $\tilde{S}$ has yet to be done.

This first validation against simulated data shows our RS theory accurately predicts the behaviour of the slope and width of the data cloud. Further, the degree of non-linearity in the assumed model affects this location i.e. the phase transition now depends on $\tilde{S}^2$. Put another way, the existence of a maximum likelihood estimator, for a given p and N, depends on the variance of the regression coefficients β^0. Informally, increasing β^0 will increase the slope of the hyperbolic tangent function in the conditional probability (3.1a) which, in turn, increases the probability of generating linearly separable data. The MLE exists only when the data is not linear separable.

Next we compare our ML results to specific results in [133] where equivalent macroscopic terms have been calculated. By switching between the two formulations (3.1a) (used in this

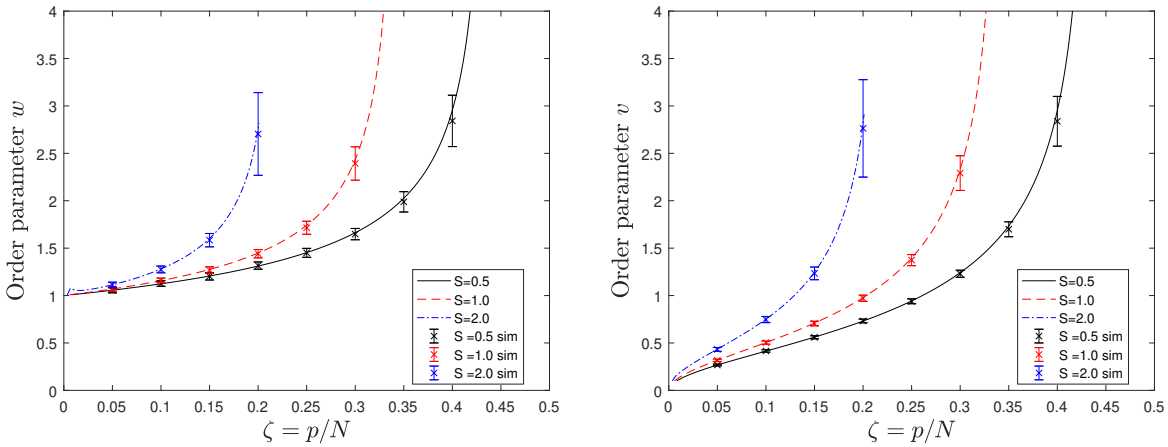

Fig. 3.5 The ML order parameters (3.32a)-(3.32c) are solved numerically to produce theoretical results (solid lines). Error bars representing one standard deviation for 100 simulations show excellent agreement with theory. We note these plots differ from [133] only in a scaling of S resulting from different formulations of the model. Recall the order parameter definitions for uncorrelated covariates $w = \lim_{p\to\infty} \frac{1}{p}\beta^0 \cdot \langle\langle\beta\rangle\rangle_{\mathscr{D}}$ and $v^2 = \lim_{p\to\infty} \frac{1}{p}\langle\langle\beta^2\rangle - \langle\beta\rangle^2\rangle_{\mathscr{D}}$.

work) and (3.1b) (used in [133]) results in rescaling the regression coefficients by a factor of one half. Also [133] scales covariates by $\sqrt{N}$ whereas our theory scales β as $\sqrt{p}$. This gives a further factor of $\sqrt{p/N} = \sqrt{\zeta}$ difference. Taking these into account[3] and using the test cases and notation in [133] with $\gamma^2 = 5$ and $\kappa \in \{0.1, 0.2\}$, we find $w = 1.1678,\ 1.4994$ versus their $\alpha_* = 1.1678,\ 1.499$ and $v = 3.3458,\ 4.7570$ versus their $\sigma_* = 3.3466,\ 4.744$. Having confirmed our theory is an accurate representation of ML inference, we proceed to investigate more general cases.

3.3.2 Uncorrelated case

Assuming uncorrelated covariates, $A = \mathbb{I}$ or $\rho(a) = \delta(a-1)$, the general equations (3.29a)-(3.29f) are now solved iteratively for $\{\tilde{u}, v, w, r, \tilde{f}, \tilde{g}\}$ in the ML and MAP cases. The results are shown in Figure 3.6 for w, v along with numerical values from synthetic data. Corresponding plots for the intercept r are included in class imbalance Section 3.3.4.

In Section 4.2.5 relating to the regularized Cox model, we consider the limit $\zeta \to \infty$. We find that $\lim_{\zeta\to\infty} v = \lim_{\zeta\to\infty} w = 0$ which corresponds to vanishing inferred association parameters with the assumed scaling of the width of the prior. This analysis is applicable to to the entire family of GLMs (see [27] for a detailed description). Empirically we found the

[3] Our raw v values were 0.52902, 1.0637. To find corresponding values from [133], we take $2v/\sqrt{\zeta}$ $v = 0.52902 * 2/\sqrt{0.1} = 3.3458,\ 1.0637 * 2/\sqrt{0.2} = 4.7570$

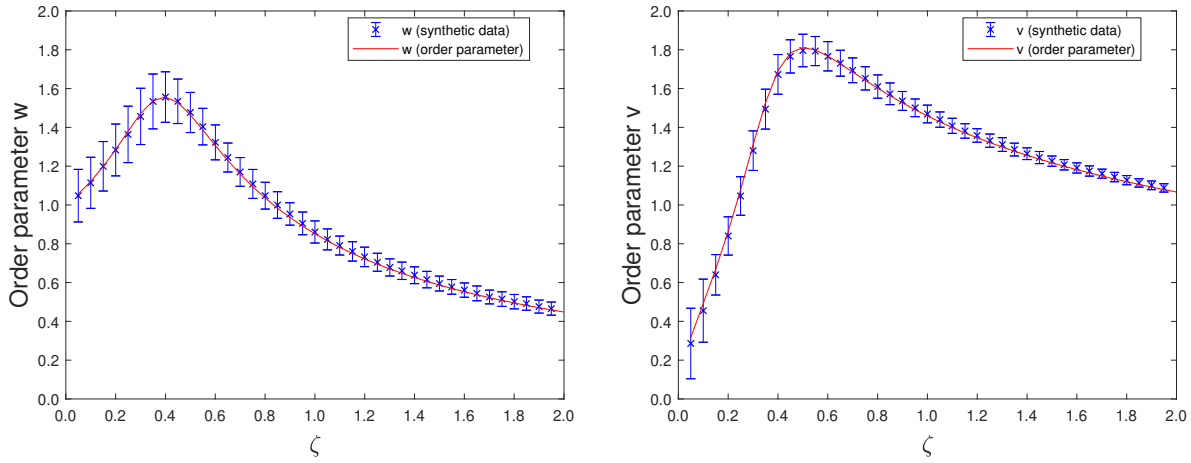

Fig. 3.6 Predicted and measured values of the order parameters w and v (solid lines and markers, respectively), for $A = \mathbb{I}$, $S = 1$ and regularization parameter $\eta = 0.025$, shown versus $\zeta = p/N \in (0,2]$. Simulations with $Np = 400,000$ are repeated 400 times with independent data sets, and results shown as averages with error bars indicating one standard deviation.

$v, w \to 0$ if numerical simulations were allowed to continue to large ζ. This means Figures 3.6 and 3.7 converge towards zero if extended ζ axis was extended further (not shown here).

Our theory predicts the relevant order parameters accurately and, as expected, the inclusion of regularization extends the regime where inference is possible beyond the ML regime. Here our replica symmetric solution appears to be valid. The important case of correlated covariates, $\mathbf{A} \neq \mathbb{I}$ is considered in the next section.

3.3.3 Correlated case

The inference of model parameters from correlated data leads to additional difficulties compared with the uncorrelated case. This effect, known as collinearity, where regression coefficients become unstable and their variances are magnified has long been known in the literature [131]. Without correct consideration, both confidence intervals and conclusions of hypothesis tests are adversely affected [44, 134] particularly as ζ increases. We defer detailed simulations and discussion of the correlated case until the next chapter.

3.3.4 Inference under class imbalance

In supervised learning tasks, training data with different class sizes leads to the minority class rarely being chosen. This so-called class imbalance problem appears to be present across classification algorithms [141]. As class imbalance increases, the intercept term

for parametrized models diverges (see [106]) leading to all new samples presented being predicted as the majority class. The intercept term is often incorporated into the regression coefficients paired with a constant covariate equal to one. However given its key role in understanding class imbalance, it has been written explicitly in our model definition (3.1a). An in-depth analysis of the role of the bias term in neural networks is contained in [142].

The importance of this problem can be understood through two quite different examples. Medical data often contains large imbalances between numbers of diseased and healthy samples. The clinically important decision is often to identify rare cases accurately. Another example is financial fraud detection where there may be millions of legitimate transactions against a handful of fraudulent ones. Again the important outcome is to identify the minority fraud class accurately. The diverging intercept term prevents accurate classification in the above examples.

Existing methods to mitigate the effect of class imbalance have focused on either data pre-processing [23, 38] or incorporating a cost function into the classification algorithm [141]. While these methods are useful to the practitioner, theoretical explanations are limited [106, 120]. Our theory, in the form of order parameter equations (3.29a)-(3.29f), contains the relevant information to investigate class imbalance effects analytically. The average imbalance is a function of the intercept parameter and can be found by defining the $\Upsilon \equiv N^{-1}\sum_{i=1}^{N} t_i \in [-1,+1]$ and using the conditional probability distribution (3.1a). Since $t \in \{-1,+1\}$, equal-sized classes results in $\Upsilon = 0$. Averaging over the data, we find

$$\begin{aligned}\langle \Upsilon \rangle_{\mathscr{D}} &= \sum_{t=\pm 1} \int_{\mathbf{z}\in\mathbb{R}^p} \mathrm{d}\mathbf{z}\; t\; p(\mathbf{z},t|\vartheta^0) = \sum_{t=\pm 1} \int \mathrm{d}\mathbf{z}\; t\; p(t|\mathbf{z},\vartheta^0) p(\mathbf{z}|\vartheta^0) \\ &= \int \mathrm{d}\mathbf{z}\;\; p(\mathbf{z})\; [p(t=1|\mathbf{z},\vartheta^0) - p(t=-1|\mathbf{z},\vartheta^0)] \\ &= \int \mathrm{d}\mathbf{z}\; p(\mathbf{z})\; \tanh\left(r^0 + \frac{\beta^0\cdot\mathbf{z}}{\sqrt{p}}\right)\end{aligned} \tag{3.34}$$

Assuming a true intercept $r^0 = 0$ results in balanced class sizes i.e. $\langle \Upsilon \rangle = 0$. Numerical integration of (3.34) for non-zero values of r^0 produces average class imbalances e.g. $r^0 = 0.5$ results in $\langle \Upsilon \rangle = 0.30$, a 35 : 65 class split.

Figure 3.7 displays our theoretical results for the intercept order parameter r. In particular, for a given level of imbalance, quantified by r^0, the amount of bias in the inferred intercept increases with ζ. Further, we find that regularization mitigates this effect leading to the possibility of correcting for class imbalanced data. This is surprising since in our theory, regularization is only applied to components of the regression coefficients $\{\beta_\mu\}_{\mu=1}^{p}$ and not to the intercept term r. It is reminiscent of the Breslow estimator for the Cox proportional

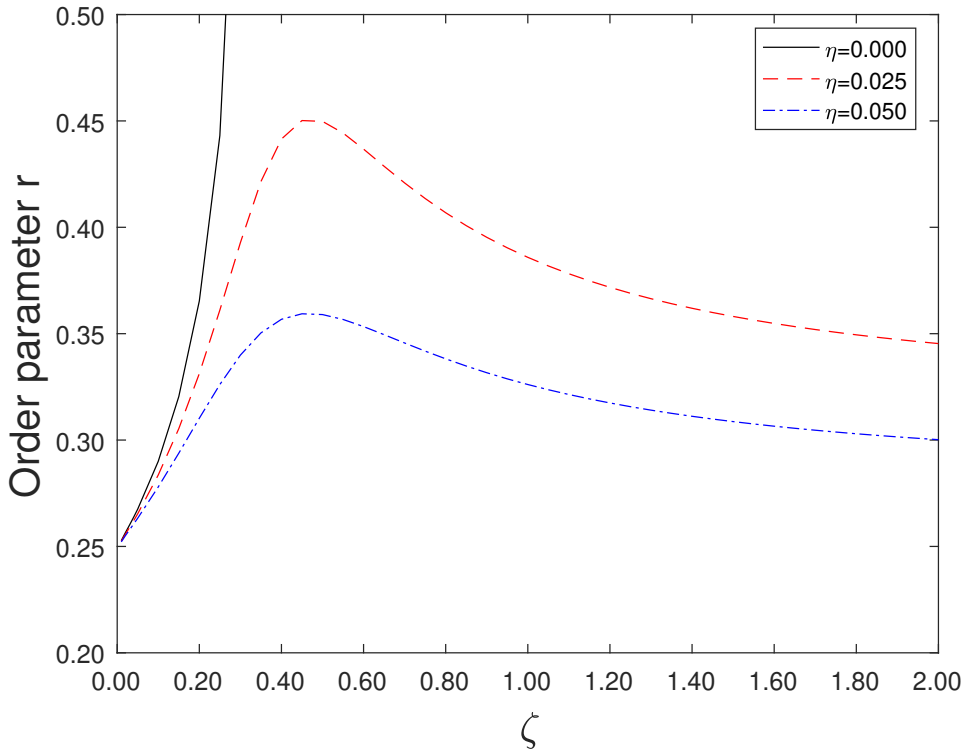

Fig. 3.7 Theoretical values of the order parameter r using $A = \mathbb{I}$, $S = 1, \eta = \{0.000, 0.025, 0.050\}$, shown versus $\zeta = p/N \in (0, 2]$. Note the truncated r-axis. The value of $r^0 = 0.25$ represents a class imbalance of $42 : 58$ calculated from (3.34). The $r^0 = 0.00$ line was calculated but not shown since it is a constant zero within numerical accuracy illustrating no theoretical bias in the intercept term when the training data has balanced class sizes.

hazards model [15] where the inferred hazard rate can be expressed as a function of the inferred regression coefficients. For the logistic regression model, it is not clear how to derive an equivalent expression for $\hat{r}$ in terms of $\hat{\beta}$. In the $\eta > 0$ case, we find the intercept inflation is not the same as that of the slope, possibly due to the different scaling factors and the regularization treatment described above. Lastly, the absence of class imbalance in the training data ($r^0 = 0$), results in accurate inference of the intercept term for all values of ζ. This can be seen by symmetry arguments outlined in appendix B.8.

3.4 Alternative method without replicas

For the linear regression model, a direct approach to calculating observable quantities was possible (see Section 2.3). We find this is no longer straightforward with the logistic regression model.

$$\mathrm{E} = \sum_{i=1}^{N} \left\{ t_i \log[1+e^{-\beta\cdot\mathbf{z}_i}] + (1-t_i)\log[1+e^{\beta\cdot\mathbf{z}_i}] \right\} + \eta\beta\cdot\beta \tag{3.35}$$

Defining $\mathbf{C} \equiv N^{-1}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i^T$ and introducing a scalar generating field λ, the required overlap function is

$$\begin{aligned}
\langle \beta^0\cdot\beta\rangle &= \lim_{\lambda\to 0}\frac{\partial}{\partial\lambda}\log\int \mathrm{d}\beta\; e^{\lambda\beta^0\cdot\beta-\gamma\sum_{i=1}^{N}\left\{t_i\log[1+e^{-\beta\cdot\mathbf{z}_i}]+(1-t_i)\log[1+e^{\beta\cdot\mathbf{z}_i}]\right\}-\gamma\eta\beta\cdot\beta} \\
&= \lim_{\lambda\to 0}\frac{\partial}{\partial\lambda}\log\int \mathrm{d}\beta\; e^{\beta\cdot(\lambda\beta^0-\gamma\eta\beta)}\prod_{i=1}^{N}[1+e^{-\beta\cdot\mathbf{z}_i}]^{-\gamma\, t_i}[1+e^{\beta\cdot\mathbf{z}_i}]^{-\gamma\,(1-t_i)} \\
&= \lim_{\lambda\to 0}\frac{\partial}{\partial\lambda}\log\int \mathrm{d}\beta\; e^{\beta\cdot(\lambda\beta^0-\gamma\eta\beta)}\prod_{i=1}^{N}e^{\frac{1}{2}\gamma\beta\cdot\mathbf{z}_i(2t_i-1)}\left(2\cosh\frac{1}{2}\beta\cdot\mathbf{z}_i\right)^{-\gamma}
\end{aligned} \tag{3.36}$$

The general β integral does not seem solvable analytically. We attempt the special case of $\gamma = 1$.

$$\langle \beta^0\cdot\beta\rangle = \lim_{\lambda\to 0}\frac{\partial}{\partial\lambda}\log\int \mathrm{d}\beta\; e^{\beta\cdot(\lambda\beta^0-\eta\beta)}\prod_{i=1}^{N}\frac{e^{\frac{1}{2}\beta\cdot\mathbf{z}_i(2t_i-1)}}{e^{\frac{1}{2}\beta\cdot\mathbf{z}_i}+e^{-\frac{1}{2}\beta\cdot\mathbf{z}_i}} \tag{3.37}$$

This does not seem feasible even in the $\gamma = 1$ case. Averaging over the true data-generating distribution does not simplify the problem and leads us towards the replica method. We conclude that the replica formalism developed in the last two chapters is indeed necessary for minimization problems for all but the most simple inference models.

3.5 Discussion

We have successfully applied our replica formalism to a Generalized Linear Model with a non-linear link function. The first surprising result is the appearance of a systematic bias in the inferred regression coefficients (see Figure 3.1) which increases with the ratio $\zeta = p/N$. This is not the case with linear regression in the previous chapter. Our theory faithfully reproduces this bias and points to a method of mitigating it via the regularization term. Interestingly, it also suggests a possible method of correcting this effect.

Separating the intercept term from other model parameters allows an investigation of the phenomenology of class imbalance leading to the next surprising result where the inferred intercept also becomes increasingly biased with ζ. A similar effect will be seen in the next chapter for the hazard rate in Cox regression.

This may lead the reader to conclude that there is a magic recipe for improving the quality of the inferred coefficients for any regression problem. Or equally magically to require less samples to be taken for the same predictive accuracy saving time and money. However our theory requires the estimation of a number of parameters $\{S^2, r^0, \mathbf{A}\}$ before being applied to real datasets. Not withstanding, this work provides valuable theoretical insight into the behaviour of macroscopic quantities relating to the regularized logistic problem with correlated covariates when $p \sim \mathcal{O}(N)$. A final problem preventing correction is the assumption of no model misspecification. For practical regression problems, the true data-generating model will be unknown leading to a degree of model mismatch.

It is important to recall that our various order parameters have been averaged over quenched disorder or true data distribution. Any particular dataset will be a single realization of the distribution. This may suggest a different approach such as the cavity method.

An analysis of class imbalance is presented in [106] for the univariate case. The behaviour of the intercept and regression coefficients is investigated with $\mathcal{O}(1)$ samples in the minority class and a diverging number in the majority class. This is labelled the infinitely imbalanced case. Replicating this regime by considering the limit $r^* \to \infty$ in our order parameter equations, not surprisingly gives $w = r = 0$ but no further useful information. This shows the limits of our asymptotic theory which is demanded by the saddle point method.

Further work. Our theory finds the difference between the true r^0 and inferred intercept values and hence the correction needed for a given level of class imbalance. This could be developed into a regularization value required for a given imbalance and p/N ratio.

The multinomial logistic regression model [105] has multiple linear predictors and a soft-max function to classify more than two classes. Our work could be extended by introducing L different regression coefficients $\{\beta_1 \cdot \mathbf{z}, \ldots, \beta_L \cdot \mathbf{z}\}$ but this significantly increases the

complexity of the algebra. Our approach could also be applied to the combination of many perceptrons into a multi-layer neural network. The conditional probability distribution of many layers of non-linear functions could be neatly represented as in [100].

Finally the link function can be changed from logit to probit or the cumulative distribution function of the Gumbel distribution allowing a wide range of binary classifiers to be analyzed.

Chapter 4

Regularized Cox model

As we found with logistic regression, when the data dimension p is comparable to the sample size N, inferred values of its regression parameters are biased due to overfitting. Unfortunately, in post-genome medicine, having large values of ζ is the rule rather than the exception. Survival analysis, commonly used to model a time-to-event random variable T, suffers this same bias prompting epidemiologists to formulate heuristic rules for avoiding overfitting, such as limits on the number of events per variable [25, 30, 63, 107, 139]. The introduction of the so-called regularized Cox model is another approach and its application to survival analysis with high-dimensional covariates is studied widely, see e.g. [71, 145] and references therein.

In the previous two chapters, statistical physics methods were applied to regression and classification problems. In this chapter, the replica formalism is used to investigate the relationship between the true and inferred regression parameters in a multivariate time-to-event[1] model - the regularized Cox proportional hazards model - in the regime of high-dimensional asymptotics. The replica analysis of [26] is generalized from ML to MAP inference, upon adding an $L2$ regularization term to the log-likelihood function.

We find, as before, that the replica symmetric version of the theory is sufficient to accurately explain the behaviour of interest. The regularization term suppresses overfitting effects, and removes the ML phase transition of the Cox model [26] at $\zeta = 1$. The resulting order parameter equations can also be used to predict the amount of regularization needed for unbiased regression, expressed in term of spectrum of $\mathbf{A}$ and the ratio ζ. This allows for straightforward overfitting corrections in time-to-event analysis in the average case.

The main mathematical difference with previous models analyzed, apart from the form of the conditional probability, is the presence of a hazard function $\lambda(t)$. It is specific to time-to-event models and can be seen as a generalization of the intercept term. Importantly

[1] The terms survival analysis and time-to-event analysis are used interchangeably in this text

$\lambda(t)$ is an unknown function rather than a scalar $r \in \mathbb{R}$ which complicates the derivation of the saddle point equations. A second difference is that the response variable t, which models the time to failure, is now a continuous non-negative random variable which results in a change from sums over a finite set to integrals over t. Before proceeding with the analysis, a brief background to survival analysis is given.

4.1 Introduction to survival analysis

Survival analysis models a continuous, non-negative random variable T representing the time until an event occurs. This could be the time to default of a company or the time to relapse of a patient. The probability density function of the random variable T is $f(t)$ and the associated cumulative distribution function

$$F(t) = P(T \leq t) = \int_0^t \mathrm{d}s\, f(s) \tag{4.1}$$

The survivor function is defined as

$$S(t) \equiv P(T > t) = 1 - F(t) \tag{4.2}$$

The hazard function is defined as the instantaneous rate of failure between t and $t + \Delta t$ given the sample has survived up to time t

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, T > t)}{\Delta t \times P(T > t)} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \times S(t)} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{dt}} \log S(t) \end{aligned} \tag{4.3}$$

From this we can find the cumulative hazard rate

$$\Lambda(t) = \int_0^t \mathrm{d}s\, \lambda(s) \tag{4.4}$$

Combining these relationships gives an alternative expression for the survivor function

$$S(t) = e^{-\int_0^t \mathrm{d}s\, \lambda(s)} = e^{-\Lambda(t)} \tag{4.5}$$

Common forms of the hazard rate. The event time T is a continuous positive real-valued random variable and can be modelled by a probability density function (pdf). Common forms

for $f(t)$ in survival analysis include the one-parameter exponential or two-parameter Weibull distributions. Both distributions have support $t \in [0, \infty)$.

$$\begin{aligned} f(t|\gamma) &= \gamma\, \mathrm{e}^{-\gamma t} && \text{Exponential} \\ f(t|\alpha,\beta) &= \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}\, \mathrm{e}^{-(t/\alpha)^{\beta}} && \text{Weibull} \end{aligned} \tag{4.6}$$

We see the exponential is a special base of the Weibull distribution with $\alpha = 1/\gamma$ and $\beta = 1$. The hazard rates corresponding to these two forms of $f(t)$ are found using (4.3). In particular, the hazard rate for the exponential function is found to be a constant value.

$$\lambda(t) = \frac{f(t)}{1-F(t)} = \frac{\gamma \mathrm{e}^{-\gamma t}}{1-[1-\mathrm{e}^{-\gamma t}]} = \gamma \tag{4.7}$$

The Weibull distribution allows for a more flexible family of hazard functions.

$$\lambda(t) = \frac{f(t)}{1-F(t)} = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-(t/\alpha)^{\beta}}}{1-[1-e^{-(t/\alpha)^{\beta}}]} = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} \tag{4.8}$$

Synthetic survival data. To make the ideas concrete, we generate 50 synthetic (non-censored) time-to-event data points and examine the resulting survivor function. The data is generated using the Weibull model (4.6) with parameters $\alpha = 250,\ \beta = 3$. This represent 50 patients in a clinical trial along with their event times T in months.

We first calculate the empirical cumulative distribution function where N is the number of data points and $\{t_i\}$ are the individual event times

$$\hat{F}_N(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{t_i<t} \tag{4.9}$$

The survivor function, $S(t) = 1 - F(t)$, is plotted as the blue staircase function in Figure 4.1. The vertical axis represents the proportion of surviving patients at time t. The survivor function at $t = 0$ is $S(t) = 1$ and $S(t = \infty) = 0$ if we assume the event will eventually occur to all patients. Each step down in the survivor function represents one or more events occurring.

The survivor function is estimated in Figure 4.1 using a parametric approach. Specifically the maximum likelihood estimator of the two Weibull parameters is calculated along with their 95% confidence intervals. There is no model mismatch since the Weibull distribution is used to both generate the data and estimate the parameters. Common non-parametric methods are the Kaplan-Meier [81] or Nelson-Aalen [1, 103] estimators. For detailed introductions to these estimators see e.g. [33, 80, 85]

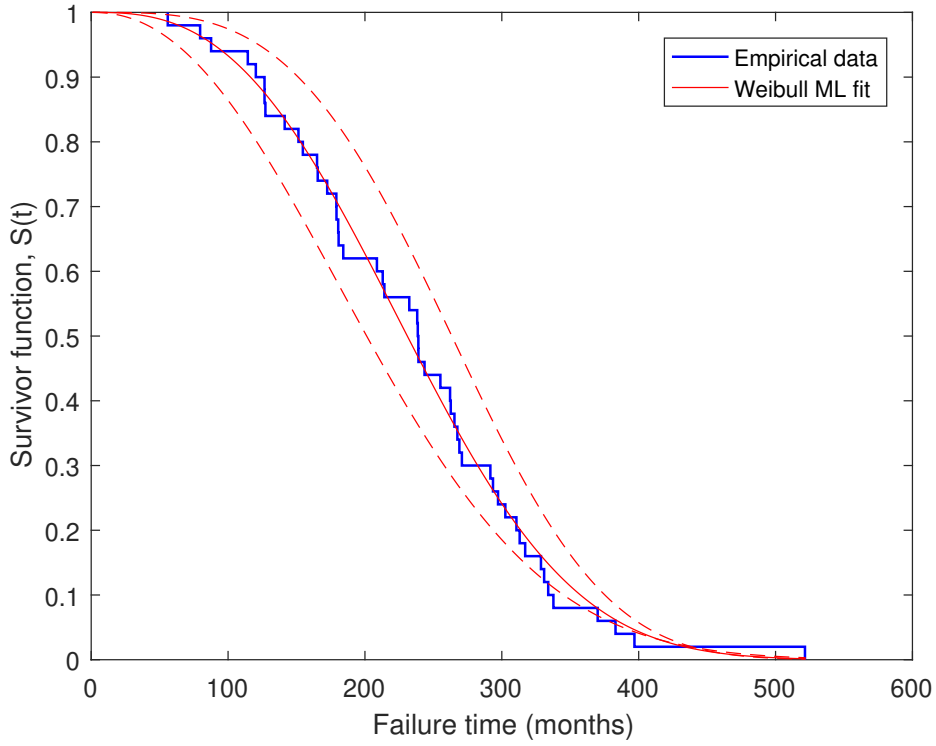

Fig. 4.1 Synthetic survival data generated with a Weibull(250,3) distribution (blue staircase function). The maximum likelihood estimator of the survivor function is shown in red along with 95% confidence intervals of the Weibull parameters (red dashed lines). This could represent, for example, 50 patients taking part in an uncensored trial with the failure time being measured in months.

In practice, time-to-event data is collected over a finite interval e.g. a clinical trial with a five year time window. This leads to the concept of censoring where survival data has the form of $(\mathbf{z}_i, t_i, \delta_i)$ detailing, for each sample i, the covariates, time-to-event and a censoring indicator ($\delta_i = 0$ if patient i does not experience the event during the fixed trial period and $\delta_i = 1$ otherwise). Patients with $\delta_i = 0$ do not provide time-to-event data but do provide valuable information that the survival time is greater than the monitoring time. This censoring can be incorporated into our current model by introducing multiple risk measures. We do not attempt this here.

Cox proportional hazards model. This model [32], commonly used in epidemiological studies and clinical trials, predicts the continuous time-to-event random variable by combining an unspecified baseline hazard rate with a function of patient covariates. The canonical form for the covariate-dependent hazard rate of this model is

$$\lambda(t|\beta, \mathbf{z}) = \lambda_0(t) e^{\beta \cdot \mathbf{z}} \tag{4.10}$$

where $\lambda_0(t)$ is the base hazard rate, $\beta \in \mathbb{R}^p$ are the regression coefficients and $\mathbf{z} \in \mathbb{R}^p$ are the (time-independent) covariates. Since $e^{\beta\cdot\mathbf{z}} = e^{\beta_1 \mathbf{z}_1} \times \ldots \times e^{\beta_1 \mathbf{z}_N}$, the hazards are "proportional".

The base hazard rate $\lambda_0(t)$ is the instantaneous probability of failure given survival to time t and all covariate values are zero. Using the specific form of the Cox model, we derive the maximum likelihood estimator of the base hazard rate by considering the functional derivative of the log likelihood with respect to the hazard rate $\lambda(t)$

$$\frac{\delta \ell}{\delta \lambda(t)} = \frac{\delta}{\delta \lambda(t)} \left\{ \log \prod_{i=1}^{N} p(t_i|\beta, \mathbf{z}_i, \lambda) \right\} = \sum_{i=1}^{N} \frac{\delta}{\delta \lambda(t)} \left\{ \log \lambda(t_i) - \Lambda(t_i) e^{\beta\cdot\mathbf{z}_i} \right\} \tag{4.11}$$

Setting $\frac{\delta \ell}{\delta \lambda(t)} = 0$ to find the ML estimate of $\lambda_0(t)$ i.e. $\mathbf{z} = (0, \ldots, 0)$ and using (B.36), we find an equivalent form to the Breslow estimator [15] of the base hazard rate.

$$\hat{\lambda}_0(t) = \frac{\sum_{i=1}^{N} \delta(t - t_i)}{\sum_{i=1}^{N} \Theta(t - t_i)} \tag{4.12}$$

The Cox proportional hazards model was originally developed for use with life-tables where N is large (population-wide data) and the number of covariates p is small. The focus is often on the so-called hazard ratios which compare the values of the factors $\exp(\beta_\mu z_\mu)$ for different covariate values. Assume A and B denote the treatment and control arms of a clinical trial. The relevant quantity is then the ratio of hazard rates

$$\frac{\lambda_A(t)}{\lambda_B(t)} = \frac{\lambda_0(t) e^{\beta\cdot\mathbf{z}_A}}{\lambda_0(t) e^{\beta\cdot\mathbf{z}_B}} = e^{\beta\cdot(\mathbf{z}_A - \mathbf{z}_B)} \tag{4.13}$$

The base hazard rates $\lambda_0(t)$ cancel hence no assumptions for the unknown $\lambda(t)$ beyond $\lambda(t) \geq 0$ are required. One can also take $z_{A\mu} = z_{B\mu}$ except for one component of $\mathbf{z}$. This leads to

$$\frac{\lambda_A(t)}{\lambda_B(t)} = e^{\beta_\mu(\mathbf{z}_{A\mu} - \mathbf{z}_{B\mu})} \tag{4.14}$$

Useful expressions specific to the Cox model are collected in Appendix B.7. There are a number of pedagogical texts [33, 80, 85] which describe alternative time-to-event models not considered in this chapter.

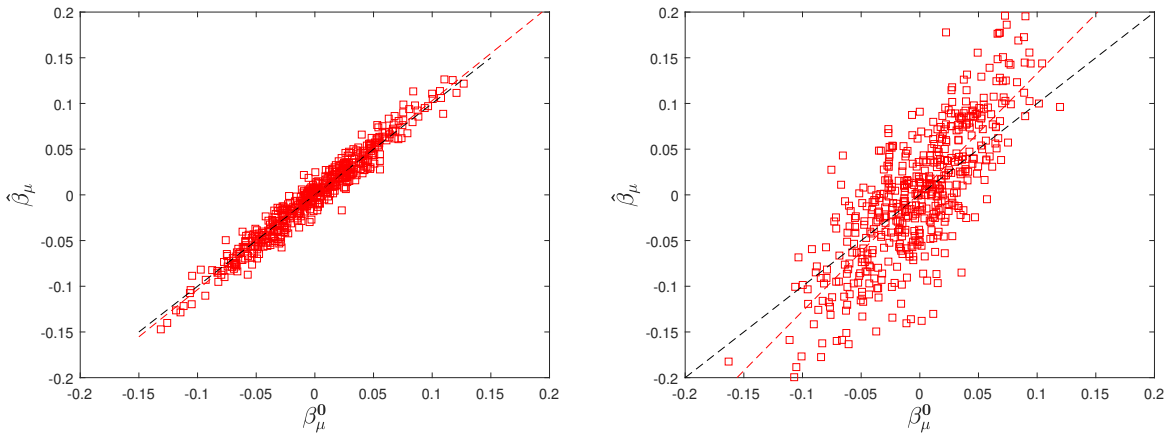

Fig. 4.2 Comparison of true β^0 and inferred $\hat{\beta}$ regression coefficients for the Cox proportional hazards model. Synthetic survival data were generated [135] using Gaussian covariates with $p = 500$ and two values of $N = 10,000$ and $N = 1,500$ resulting in $\zeta = 0.05$ (left) and $\zeta = 0.33$ (right). The red dashed line is the best fit to the data cloud and the black dashed line $\hat{\beta} = \beta^0$ represents perfect inference. The regression coefficients are estimated via maximum likelihood inference (using the R package *glmnet* [56]). Both the slope and variance of the data cloud increases with ζ.

4.2 Replica analysis of regularized Cox regression

Once again, we start with the maximum a posteriori form of the overfitting measure

$$\begin{aligned} \mathrm{E}(\vartheta^0, \mathscr{D}) &\equiv \min_{\vartheta} \left\{ D(\hat{P}_{\mathscr{D}} \| P_{\vartheta}) - \log p(\vartheta) \right\} - \left\{ D(\hat{P}_{\mathscr{D}} \| P_{\vartheta^0}) - \log p(\vartheta^0) \right\} \\ &= \min_{\vartheta} \left\{ \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(t_i|\mathbf{z}_i, \vartheta^0)\, p(\vartheta^0)}{p(t_i|\mathbf{z}_i, \vartheta)\, p(\vartheta)} \right\} \end{aligned} \tag{4.15}$$

Following the methods of the previous chapters, we interpret minimization of (4.15) as computing the ground state energy of a statistical mechanical system with degrees of freedom ϑ and Hamiltonian $H(\vartheta|\vartheta^0, \mathscr{D})$, at inverse temperature γ, where $\mathscr{D} = \{(t_1, \mathbf{z}_1), \ldots, (t_N, \mathbf{z}_N)\}$ and

$$H(\vartheta|\vartheta^0, \mathscr{D}) = \log \prod_{i=1}^{N} \left[\frac{p(t_i|\mathbf{z}_i, \vartheta^0)\ p(\vartheta^0)}{p(t_i|\mathbf{z}_i, \vartheta)\ p(\vartheta)} \right] \tag{4.16}$$

We define the associated free energy, which we average over the disorder (the microscopic realization of $\mathscr{D}$), and can compute the disorder-averaged ground state energy as the $\gamma \to \infty$ limit of the disorder-averaged energy density $\mathrm{E}_\gamma(\vartheta^0)$, where

$$\mathrm{E}_\gamma(\vartheta^0) = -\frac{1}{N} \frac{\partial}{\partial \gamma} \left\langle \log \int \mathrm{d}\vartheta \prod_{i=1}^{N} \left[\frac{p(t_i|\mathbf{z}_i, \vartheta) p(\vartheta)}{p(t_i|\mathbf{z}_i, \vartheta^0) p(\vartheta^0)} \right]^\gamma \right\rangle_{\mathscr{D}} \tag{4.17}$$

The replica identity $\langle \log Z \rangle = \lim_{n\to 0} n^{-1} \log \langle Z^n \rangle$ is subsequently used to simplify the average of the logarithm, giving in the present case

$$\begin{aligned}
\mathrm{E}_\gamma(\vartheta^0) &= -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \left\langle \left\{ \int \mathrm{d}\vartheta \prod_{i=1}^N \left[\frac{p(t_i|\mathbf{z}_i,\vartheta) p(\vartheta)}{p(t_i|\mathbf{z}_i,\vartheta^0) p(\vartheta^0)} \right]^\gamma \right\}^n \right\rangle_{\mathscr{D}} \\
&= -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \Big\{ \prod_{\alpha=1}^n \mathrm{d}\vartheta^\alpha \Big[\frac{p(\vartheta^\alpha)}{p(\vartheta^0)} \Big]^\gamma \Big\} \Big\langle \prod_{i=1}^N \prod_{\alpha=1}^n \Big[\frac{p(t_i|\mathbf{z}_i,\vartheta^\alpha))}{p(t_i|\mathbf{z}_i,\vartheta^0)} \Big]^\gamma \Big\rangle_{\mathscr{D}} \\
&= -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \Big\{ \prod_{\alpha=1}^n \mathrm{d}\vartheta^\alpha \Big[\frac{p(\vartheta^\alpha)}{p(\vartheta^0)} \Big]^\gamma \Big\} \\
&\times \Big\{ \int \mathrm{d}\mathbf{z}\mathrm{d}t \; p(\mathbf{z}) p(t|\mathbf{z},\vartheta^0) \prod_{\alpha=1}^n \Big[\frac{p(t|\mathbf{z},\vartheta^\alpha))}{p(t|\mathbf{z},\vartheta^0)} \Big]^\gamma \Big\}^N
\end{aligned} \tag{4.18}$$

We will now make a specific choice for $p(t|\mathbf{z},\vartheta)$, and use (4.18) to develop a theory for regression and overfitting in regularized Cox models with Gaussian priors.

4.2.1 Application to the regularized Cox proportional hazards model

Using (4.3)-(4.5) along with the definition of the Cox proportional hazards model (4.10) originally described in [32] takes the following form

$$p(t|\mathbf{z},\vartheta) = \lambda_0(t) \mathrm{e}^{\beta\cdot\mathbf{z} - \exp(\beta\cdot\mathbf{z}) \int_0^t \mathrm{d}t' \; \lambda_0(t')} \tag{4.19}$$

The response variable represents the time-to-event $t \in \mathbb{R}^+$ rather than $t \in \mathbb{R}$ or $t \in \{-1,+1\}$. The hazard rate $\lambda(t)$ is a non-negative function defined for $0 \leq t < \infty$ which can be thought of as the logarithm of the intercept term in the previous two models. The model parameters to be inferred are $\beta \in \mathbb{R}^p$ and the hazard rate $\lambda(t)$. Substituting $\vartheta = \{\beta, \lambda\}$ translates (4.18) into

$$\begin{aligned}
\mathrm{E}_\gamma(\beta^0,\lambda^0) = -\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \{\mathrm{d}\lambda^1 \ldots \mathrm{d}\lambda^n\} \int \mathrm{d}\beta^1 \ldots \mathrm{d}\beta^n \Big\{ \prod_{\alpha=1}^n \Big[\frac{p(\beta^\alpha)}{p(\beta^0)} \Big]^\gamma \Big\} \\
\times \Big\{ \int \mathrm{d}\mathbf{z}\mathrm{d}t \; p(\mathbf{z}) p(t|\mathbf{z},\beta^0,\lambda^0) \prod_{\alpha=1}^n \Big[\frac{p(t|\mathbf{z},\beta^\alpha,\lambda^\alpha))}{p(t|\mathbf{z},\beta^0,\lambda^0)} \Big]^\gamma \Big\}^N
\end{aligned} \tag{4.20}$$

Functional integrals are written as $\int \{\mathrm{d}\lambda\}$, the true model parameters are $\{\beta^0,\lambda^0\}$, and we follow the standard convention for regularized Cox models of only including a prior for the association parameters (equivalently, assuming an improper, or 'flat', prior for the base hazard rate). Our *L2* prior is $p(\beta) \propto \exp(-p\eta\beta^2)$, and we will find in our analysis that this form indeed gives the appropriate scaling with p. To proceed with the analytical treatment, we

assume that the covariate vectors $\mathbf{z}_i$ are drawn independently from a population distribution with zero mean and covariance matrix $\mathbf{A}$. Our analysis is carried out in the regime where both $N, p \to \infty$ but with fixed ratio $\zeta = p/N \sim \mathscr{O}(1)$. To retain non-zero event times, even for $p \to \infty$, we must rescale the regression coefficients according to $\beta \to \beta/\sqrt{p}$, resulting in $\beta \cdot \mathbf{z} \sim \mathscr{O}(1)$. Without this rescaling we would have event time distributions with all weight concentrated on $t \to 0$ and $t \to \infty$. We follow the methodology of the previous two chapters but show relevant milestones to fix the slightly different notation in the reader's mind.

Starting from (4.20), we introduce the Dirac delta function to transport the linear predictor $\beta \cdot \mathbf{z}$ out of the main integral, convert the expression to the saddle point integral form and assume replica symmetry to arrive at

$$\lim_{N\to\infty} \mathrm{E}_\gamma(\beta^0, \lambda^0) = \frac{\partial}{\partial \gamma} \mathrm{extr}_{u,v,w,f,g,\lambda} \Psi_{\mathrm{RS}}(u,v,w,f,g,\lambda) \tag{4.21}$$

in which

$$\begin{aligned}
\Psi_{\mathrm{RS}}(\ldots) = &-\frac{1}{2}\zeta(g+f)u^2 - \frac{1}{2}\zeta g(v^2+w^2) - \zeta\eta\gamma S^2 \\
&+ \frac{1}{2}\zeta\left\{\tilde{S}^2 w^2 \left\langle \frac{a^2(\beta^0\cdot\mathbf{v})^2}{2\eta\gamma+ga}\right\rangle^{-1} + \left\langle \log\left(\frac{2\eta\gamma+ga}{a}\right)\right\rangle + f\left\langle \frac{a}{2\eta\gamma+ga}\right\rangle\right\} \\
&- \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|\tilde{S}y_0, \lambda^0) \log \int \mathrm{D}y\ \frac{p^\gamma(t|uy+wy_0+vz,\lambda)}{p^\gamma(t|\tilde{S}y_0,\lambda^0)}
\end{aligned} \tag{4.22}$$

Once again, we notice this expression is split into a model independent Ψ^A_{RS} and the model specific term Ψ^B_{RS} as with (2.52). Our replica symmetric theory thereby becomes

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}_\gamma(\beta^0, \lambda^0) = &\int \mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) \log p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) - \zeta\eta S^2 \\
&+ \eta\zeta\left\{ w^2\langle a\rangle \left\langle \frac{a^2}{2\eta\gamma+ga}\right\rangle^{-2} \left\langle \frac{a^2}{(2\eta\gamma+ga)^2}\right\rangle + \left\langle \frac{1}{2\eta\gamma+ga}\right\rangle - f\left\langle \frac{a}{(2\eta\gamma+ga)^2}\right\rangle\right\} \\
&- \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) \frac{\int \mathrm{D}y\ p^\gamma(t|uy+wy_0+vz,\lambda) \log p(t|uy+wy_0+vz,\lambda)}{\int \mathrm{D}y\ p^\gamma(t|uy+wy_0+vz,\lambda)}
\end{aligned} \tag{4.23}$$

The scalar order parameters $(u,v,w,f,g\}$ and the function $\lambda(t)$ are computed by extremization of the following function, from which we removed any constant terms:

$$\begin{aligned}\Psi_{\rm RS}(\ldots) &= -\frac{1}{2}\zeta(g+f)u^2 - \frac{1}{2}\zeta g(v^2+w^2) \\ &+ \frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta\gamma+ga}\right\rangle^{-1} + \left\langle\log(2\eta\gamma+ga)\right\rangle + f\left\langle\frac{a}{2\eta\gamma+ga}\right\rangle\right\} \\ &- \int {\rm D}z{\rm D}y_0 \int {\rm d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\log\int {\rm D}y\ p^\gamma(t|uy+wy_0+vz,\lambda)\end{aligned} \tag{4.24}$$

In order to take the the limit $\gamma\to\infty$, we once again use the saddle point integral by defining $q\equiv y/\sqrt{\gamma}$

$$\begin{aligned}\log\int Dy\ p^\gamma(t|uy+wy_0+vz,r) &= \log\int\frac{{\rm d}y}{\sqrt{2\pi}}\ \exp\left[-\frac{1}{2}y^2+\gamma\ \log p(t|uy+wy_0+vz,r)\right] \\ &= \log\int\frac{{\rm d}q\sqrt{\gamma}}{\sqrt{2\pi}}\ \exp\gamma\left[-\frac{1}{2}q^2+\log p(t|\tilde{u}q+wy_0+vz,r)\right]\end{aligned} \tag{4.25}$$

4.2.2 Scaling of order parameters with γ

We will only be interested in the limit $\gamma\to\infty$, where the stochastic process becomes deterministic MAP inference. Once again we make the following ansatz for the scaling with γ of the scalar order parameters:

$$u=\tilde{u}/\sqrt{\gamma},\quad v,w=\mathscr{O}(1),\quad g=\tilde{g}\gamma,\quad f=\tilde{f}\gamma^2 \tag{4.26}$$

Insertion into (4.24), followed by taking the limit $\gamma\to\infty$, gives

$$\begin{aligned}\lim_{\gamma\to\infty}\frac{1}{\gamma}\Psi_{\rm RS}(\ldots) &= \frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}+\tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle-\tilde{u}^2\right]-\tilde{g}(v^2+w^2)\right\} \\ &- \int {\rm D}z{\rm D}y_0\int {\rm d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\lim_{\gamma\to\infty}\frac{1}{\gamma}\log\int {\rm d}y\ {\rm e}^{\gamma\left[\log p(t|\tilde{u}y+wy_0+vz,\lambda)-\frac{1}{2}y^2\right]} \\ &= \frac{1}{2}\zeta\left\{w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1}+\tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle-\tilde{u}^2\right]-\tilde{g}(v^2+w^2)\right\} \\ &- \int {\rm D}z{\rm D}y_0\int {\rm d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\max_y\left[\log p(t|\tilde{u}y+wy_0+vz,\lambda)-\frac{1}{2}y^2\right]\end{aligned} \tag{4.27}$$

Up to this point, the replica calculations have been similar to the previous two chapters. In fact, (4.27) is almost identical to (2.69) and (3.8) apart from the functional form of the hazard rate λ compared with the scalar intercept term r. The maximization over y with our specific conditional probability will lead to differences in the order parameter equations.

We define $\phi(q) \equiv \log p(t|\tilde{u}q+wy_0+vz,\lambda) - \frac{1}{2}q^2$, recall the conditional probability (B.33) and compute $\frac{\partial\phi(q)}{\partial q} = 0$: $q = \tilde{u} - \tilde{u}e^{\tilde{u}q+\eta}\Lambda_0(t)$ with the shorthand $\eta = wy_0+vz$. Using the transformation $q = \tilde{u} - \frac{x}{\tilde{u}}$

$$\begin{aligned} &-(\tilde{u}-\frac{x}{\tilde{u}}) + \tilde{u} - \tilde{u}e^{\tilde{u}(\tilde{u}-\frac{x}{\tilde{u}})+\eta}\Lambda_0(t) = 0 \\ &\Rightarrow x = \tilde{u}^2 e^{\tilde{u}^2+\eta}e^{-x}\Lambda_0(t) \\ &\Rightarrow xe^x = \tilde{u}^2 e^{\tilde{u}^2+\eta}\Lambda_0(t) \\ &\Rightarrow x = W\left(\tilde{u}^2 e^{\tilde{u}^2+\eta}\Lambda_0(t)\right) \end{aligned} \tag{4.28}$$

in which $W(x)$ denotes Lambert's W-function, i.e. the inverse of $f(x) = xe^x$. Useful identities are found in Appendix A.6.

$$\begin{aligned} \operatorname{argmax}_y\left[\log p(t|\tilde{u}y+wy_0+vz,\lambda) - \frac{1}{2}y^2\right] &= \tilde{u} - \frac{1}{\tilde{u}}W\left(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\right) \\ \max_y\left[\log p(t|\tilde{u}y+wy_0+vz,\lambda) - \frac{1}{2}y^2\right] &= \frac{1}{2}(\tilde{u}^2+\tilde{u}^{-2}) + wy_0 + vz + \log\lambda(t) \\ &\quad - \frac{1}{2\tilde{u}^2}\left[W\left(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\right)+1\right]^2 \end{aligned} \tag{4.29}$$

This then results in

$$\begin{aligned} \lim_{\gamma\to\infty}\frac{1}{\gamma}\Psi_{\mathrm{RS}}(\ldots) &= \frac{1}{2}\zeta\left[w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-1} + \tilde{f}\left[\left\langle\frac{a}{2\eta+\tilde{g}a}\right\rangle - \tilde{u}^2\right] - \tilde{g}(v^2+w^2)\right] \\ &\quad + \frac{1}{2\tilde{u}^2}\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\left[W\left(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\right)+1\right]^2 \\ &\quad - \frac{1}{2}(\tilde{u}^2+\tilde{u}^{-2}) - \int \mathrm{D}y_0\int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\log\lambda(t) \end{aligned} \tag{4.30}$$

Similarly, working out (4.23) in the limit $\gamma\to\infty$ gives

$$\begin{aligned} \lim_{N\to\infty}\mathrm{E}_\infty(\beta^0,\lambda^0) &= \int \mathrm{D}y_0\int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\left[\log p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0) - \log\lambda(t)\right] \\ &\quad - \tilde{u}^2 - \zeta\eta S^2 + \eta\zeta\left[w^2\langle a\rangle\left\langle\frac{a^2}{2\eta+\tilde{g}a}\right\rangle^{-2}\left\langle\frac{a^2}{(2\eta+\tilde{g}a)^2}\right\rangle - \tilde{f}\left\langle\frac{a}{(2\eta+\tilde{g}a)^2}\right\rangle\right] \\ &\quad + (1+\tilde{u}^{-2})\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)W\left(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\right) \end{aligned} \tag{4.31}$$

What remains in our RS analysis is to determine the order parameters $\{\tilde{u}, v, w, \tilde{f}, \tilde{g}, \lambda\}$ by extremization of (4.30), and to substitute the result into (4.31).

4.2.3 Scalar saddle point equations

Partial differentiation of (4.30) with respect to the five scalar order parameters $\{\tilde{u}, v, w, \tilde{f}, \tilde{g}\}$ is now straightforward and gives, upon using identities such as $W'(z) = W(z)/z[1+W(z)]$ and some manipulations

$$\zeta \tilde{f}\tilde{u}^4 = -\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) \Big[W\Big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2 + w y_0 + v z} \Lambda(t)\Big) - \tilde{u}^2 \Big]^2 \tag{4.32a}$$

$$\zeta \tilde{g}\tilde{u}^2 = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) \frac{W\big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2 + w y_0 + v z} \Lambda(t)\big)}{1 + W\big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2 + w y_0 + v z} \Lambda(t)\big)} \tag{4.32b}$$

$$0 = \zeta w \Big[\langle a \rangle \Big\langle \frac{a^2}{2\eta + \tilde{g} a} \Big\rangle^{-1} - \tilde{g} \Big] \tag{4.32c}$$

$$+ \frac{1}{\tilde{u}^2} \int \mathrm{D}z\mathrm{D}y_0\ y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) W\Big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2 + w y_0 + v z} \Lambda(t)\Big) \tag{4.32d}$$

$$\tilde{u}^2 = \Big\langle \frac{a}{2\eta + \tilde{g} a} \Big\rangle \tag{4.32e}$$

$$v^2 = w^2 \Big[\langle a \rangle \Big\langle \frac{a^2}{2\eta + \tilde{g} a} \Big\rangle^{-2} \Big\langle \frac{a^3}{(2\eta + \tilde{g} a)^2} \Big\rangle - 1 \Big] - \tilde{f} \Big\langle \frac{a^2}{(2\eta + \tilde{g} a)^2} \Big\rangle \tag{4.32f}$$

Compared to the simpler scenario of [26], the present RS theory involves two additional order parameters, $\tilde{f}$ and $\tilde{g}$. Reversing the order of differentiation with respect to order parameters and the limit $\gamma \to \infty$ results in identical saddle point equations.

4.2.4 Functional saddle point equation

The equation from which to solve the functional order parameter $\lambda(t)$ is derived by functional differentiation of (4.30). Upon using the short-hand $p(t) = \int \mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0)$ for the typical distribution of the event times in the data, this equation takes the form

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y_0 \int_t^\infty \frac{\mathrm{d}t'}{\tilde{u}^2 \Lambda(t')}\ p(t'|S\langle a\rangle^{\frac{1}{2}} y_0, \lambda^0) W\Big(\tilde{u}^2 \mathrm{e}^{\tilde{u}^2 + w y_0 + v z} \Lambda(t')\Big) \tag{4.33}$$

It differs only minimally from the one in [26], and is hence equally difficult to solve analytically. We will therefore follow a variational approach, motivated by the asymptotic form of the solution for large times (see [26] for details), and choose the functional ansatz

$$\Lambda(t) = k[\Lambda^0(t)]^\rho \tag{4.34}$$

Now only two variational parameters (k,ρ) are to be estimated, rather than a function. Inserting (4.34) into (4.30), followed by partial differentiation with respect to k and ρ then leads to the following two equations:

$$\tilde{u}^2 = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0) W\Big(k\tilde{u}^2 \mathrm{e}^{\tilde{u}^2+wy_0+vz}[\Lambda^0(t)]^\rho\Big) \tag{4.35a}$$

$$\begin{aligned} 0 = {} & \frac{1}{\tilde{u}^2}\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0) W\Big(k\tilde{u}^2 \mathrm{e}^{\tilde{u}^2+wy_0+vz}[\Lambda^0(t)]^\rho\Big)\log\Lambda^0(t) \\ & -\frac{1}{\rho} - \int \mathrm{D}y_0 \int \mathrm{d}t\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\log\Lambda^0(t) \end{aligned} \tag{4.35b}$$

These have to be solved numerically alongside (4.32a)-(4.32f). We will compactify our equations by using instead of k the variable $q = k\tilde{u}^2\exp(\tilde{u}^2)$. As a further benefit of our variational ansatz, the time integrations in the saddle point equations can be simplified significantly upon switching to the new integration variable $s = \exp[-\exp(S\langle a\rangle^{\frac{1}{2}}y_0)\Lambda^0(t)] \in [0,1]$, which gives $\mathrm{d}s = -p(t|S\langle a\rangle^{\frac{1}{2}}y_0,\lambda^0)\mathrm{d}t$. Noting that $\{y_0,z\}$ always occur together allows us to combine them into a single Gaussian variable x with zero mean and variance $\sigma^2 = (w-\rho S\langle a\rangle^{\frac{1}{2}})^2 + v^2$, giving

$$\zeta\tilde{f}\tilde{u}^4 = -\int \mathrm{D}x \int_0^1 \mathrm{d}s\ \Big[W\big(q\mathrm{e}^{\sigma x}\log^\rho(1/s)\big) - \tilde{u}^2\Big]^2 \tag{4.36a}$$

$$\zeta\tilde{g}\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s\ \frac{W\big(q\mathrm{e}^{\sigma x}\log^\rho(1/s)\big)}{1+W\big(q\mathrm{e}^{\sigma x}\log^\rho(1/s)\big)} \tag{4.36b}$$

$$w = \frac{\tilde{g}\rho S}{\langle a\rangle^{\frac{1}{2}}}\Big\langle \frac{a^2}{2\eta+\tilde{g}a}\Big\rangle \tag{4.36c}$$

$$\tilde{u}^2 = \Big\langle \frac{a}{2\eta+\tilde{g}a}\Big\rangle \tag{4.36d}$$

$$v^2 = w^2\Big[\langle a\rangle\Big\langle \frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-2}\Big\langle \frac{a^3}{(2\eta+\tilde{g}a)^2}\Big\rangle - 1\Big] - \tilde{f}\Big\langle \frac{a^2}{(2\eta+\tilde{g}a)^2}\Big\rangle \tag{4.36e}$$

$$\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s\ W\big(q\mathrm{e}^{\sigma x}\log^\rho(1/s)\big) \tag{4.36f}$$

$$\frac{\tilde{u}^2}{\rho} = \int \mathrm{D}x \int_0^1 \mathrm{d}s\ W\Big(q\mathrm{e}^{\sigma x}\log^\rho(\frac{1}{s})\Big)\log\log(\frac{1}{s}) - \zeta\tilde{g}\tilde{u}^2 S\langle a\rangle^{\frac{1}{2}}(w-\rho S\langle a\rangle^{\frac{1}{2}}) + \tilde{u}^2 C_{\mathrm{E}} \tag{4.36g}$$

in which C_{E} denotes Euler's constant, and where we used the integral $\int_0^1 \mathrm{d}s\ \log\log(1/s) = \int_0^\infty \mathrm{d}x\ \mathrm{e}^{-x}\log x = -C_{\mathrm{E}}$.

4.2.5 The limits $\eta \to 0$, $\zeta \to 0$ and $\zeta \to \infty$

Here we investigate the order parameter behaviour in the small and large ζ limits. This allows us to confirm the shape of the order parameter plots, both analytically and by numerical analysis. In particular, since the order parameters v and w increase with ζ for small ζ and tend to zero for large ζ, we conclude that there must be a stationary point between these two extremes. This analytical argument is validated by numerical solutions of (4.36a)-(4.36g) (see Figure 4.4) and by simulations. An explanation of this phenomenon in terms of model complexity and the emergence of statistical constraints can be constructed [18]. In the limit $\eta \to 0$, describing a fully flat prior for association parameters, the regression changes from MAP to ML, and our RS equations should therefore reduce to those of [26]. Upon setting $\eta \to 0$ in (4.36a)–(4.36g), we immediately find that

$$w = \rho S\langle a\rangle^{\frac{1}{2}}, \qquad \tilde{g} = 1/\tilde{u}^2, \qquad \tilde{f} = -v^2/\tilde{u}^4 \tag{4.37}$$

From the first of these it follows that $\sigma = v$, and that the remaining RS scalar order parameter equations from which to solve $\{v, \rho, \tilde{u}, q\}$ hence simplify to

$$\zeta v^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s \left[W\left(q\mathrm{e}^{vx}\log^\rho(1/s)\right) - \tilde{u}^2\right]^2 \tag{4.38a}$$

$$\zeta = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, \frac{W\left(q\mathrm{e}^{vx}\log^\rho(1/s)\right)}{1+W\left(q\mathrm{e}^{vx}\log^\rho(1/s)\right)} \tag{4.38b}$$

$$\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q\mathrm{e}^{vx}\log^\rho(1/s)\right) \tag{4.38c}$$

$$\frac{\tilde{u}^2}{\rho} = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q\mathrm{e}^{vx}\log^\rho(\frac{1}{s})\right)\log\log(\frac{1}{s}) + \tilde{u}^2 C_{\mathrm{E}} \tag{4.38d}$$

For $\zeta \to 0$ (no overfitting) we expect to find $v \to 0$ and $w, \rho, k \to 1$. In analogy with [26] we now make the ansätze that $\tilde{u}, v = O(\sqrt{\zeta})$ and $\rho = 1 + O(\zeta)$ for $\zeta \to 0$, and expand (4.36a)–(4.36g) in leading order for small ζ, using $W(z) = z + O(z^2)$ for $z \to 0$. After expanding the various integrals, whose leading orders in ζ can all be done analytically, this results in

$$\begin{aligned} &\tilde{u}^2/\zeta = 1 + O(\zeta), \qquad \zeta\tilde{g} = 1 + O(\zeta), \qquad w = S\langle a\rangle^{\frac{1}{2}} + O(\zeta), \\ &\tilde{f}\zeta = -1 + O(\zeta), \qquad k = 1 + O(\zeta), \qquad v^2/\zeta = 1 + O(\zeta), \end{aligned} \tag{4.39}$$

which confirms that, in the absence of overfitting, we indeed recover the correct values of the order parameters from our RS equations.

Finally we inspect the behaviour of the RS equations (4.36a)–(4.36g) in the limit $\zeta \to \infty$ of a diverging imbalance between the number of covariates and the number of samples. This limit is inaccessible in the ML case due to a phase transition at $\zeta = 1$, where $v, w \to \infty$. In the present theory, describing the regularized version of the Cox model, this phase transition is suppressed by the Bayesian prior, provided we choose $\eta > 0$. We now make the ansatz that $\tilde{g} \to 0$ for $\zeta \to \infty$, giving $\tilde{u}^2 \to \langle a \rangle / 2\eta$, $v \to 0$, $w \to 0$, $\tilde{f} \to 0$, $\sigma^2 \to \rho^2 S^2 \langle a \rangle$, and upon introducing $Q = \lim_{\zeta \to \infty} \zeta \tilde{g} \tilde{u}^2$, the remaining trio $\{Q, q, \rho\}$ is for $\zeta \to \infty$ to be solved from the remaining three coupled equations

$$Q = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, \frac{W\left(q \mathrm{e}^{\rho S \langle a \rangle^{\frac{1}{2}} x} \log^{\rho}(1/s)\right)}{1 + W\left(q \mathrm{e}^{\sigma x} \log^{\rho}(1/s)\right)} \tag{4.40a}$$

$$\frac{\langle a \rangle}{2\eta} = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q \mathrm{e}^{\rho S \langle a \rangle^{\frac{1}{2}} x} \log^{\rho}(1/s)\right) \tag{4.40b}$$

$$\frac{\langle a \rangle}{2\eta\rho} = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\left(q \mathrm{e}^{\rho S \langle a \rangle^{\frac{1}{2}} x} \log^{\rho}(\frac{1}{s})\right) \log\log(\frac{1}{s}) + QS^2 \langle a \rangle \rho + \frac{\langle a \rangle}{2\eta} C_{\mathrm{E}} \tag{4.40c}$$

For $\zeta \to \infty$ we thus expect to find, as a consequence of $\lim_{\zeta \to \infty} v = \lim_{\zeta \to \infty} w = 0$, vanishing inferred association parameters in the present regularized Cox model, with the assumed scaling of the width of the prior.

4.2.6 Expression for the overfitting measure

Finally, using the variational approximation for the cumulative hazard rate, the manipulations applied to the RS saddle point equations, and the actual order parameter equations themselves, the overfitting measure (4.31) can be simplified to the form

$$\begin{aligned} \lim_{N \to \infty} \mathrm{E}_{\infty}(\beta^0, \lambda^0) = \eta\zeta \Big[w^2 \langle a \rangle \Big\langle \frac{a^2}{2\eta + \tilde{g}a} \Big\rangle^{-2} \Big\langle \frac{a^2}{(2\eta + \tilde{g}a)^2} \Big\rangle - \tilde{f} \Big\langle \frac{a}{(2\eta + \tilde{g}a)^2} \Big\rangle \Big] \\ + \int \mathrm{d}t \; p(t) \log\Big(\frac{\lambda^0(t)}{\lambda(t)}\Big) - \zeta\eta S^2 \end{aligned} \tag{4.41}$$

with the short-hand $p(t) = \int \mathrm{D}y_0\ p(t|S\langle a\rangle^{\frac{1}{2}}y_0, \lambda^0)$. Our variational ansatz $\Lambda(t) = k[\Lambda^0(t)]^\rho$ implies that $\lambda(t) = k\rho\lambda^0(t)[\Lambda^0(t)]^{\rho-1}$, hence

$$\begin{aligned}\int \mathrm{d}t\ p(t)\log\Big(\frac{\lambda^0(t)}{\lambda(t)}\Big) &= -\log k - \log\rho - (\rho-1)\int \mathrm{d}t\ p(t)\log\Lambda^0(t)\\ &= -\log k - \log\rho - (\rho-1)\int \mathrm{D}y_0 \int_0^1 \mathrm{d}s\Big[\mathrm{e}^{-S\langle a\rangle^{\frac{1}{2}}y_0}\log(\frac{1}{s})\Big]\\ &= -\log k - \log\rho - (\rho-1)\int_0^\infty \mathrm{d}x\ \mathrm{e}^{-x}\log x\\ &= -\log k - \log\rho + (\rho-1)C_{\mathrm{E}}\end{aligned} \tag{4.42}$$

Our final result for the asymptotic overfitting measure $\mathrm{E}(S) = \lim_{N\to\infty}\mathrm{E}_\infty(\beta^0, \lambda^0)$ is therefore

$$\begin{aligned}\mathrm{E}(S) = \eta\zeta\Big[w^2\langle a\rangle\Big\langle\frac{a^2}{2\eta+\tilde{g}a}\Big\rangle^{-2}\Big\langle\frac{a^2}{(2\eta+\tilde{g}a)^2}\Big\rangle - \tilde{f}\Big\langle\frac{a}{(2\eta+\tilde{g}a)^2}\Big\rangle\Big]\\ -\log k - \log\rho + (\rho-1)C_{\mathrm{E}} - \zeta\eta S^2\end{aligned} \tag{4.43}$$

We observe that, as was the case in [26] (without regularization), both the RS order parameter equations and the overfitting measure have within the variational approximation become completely independent of the true base hazard rate $\lambda^0(t)$.

4.3 Numerical experiments

4.3.1 Covariance distributions

Covariates of real survival data can be distributed in many different ways. The assumption originally outlined in (2.29) of Gaussian distributed risk scores is a direct consequence of working in the limit $p\to\infty$, in combination with the Central Limit Theorem. More specifically, there is no need to assume Gaussian covariate statistics. To verify the validity of Gaussian risk score statistics, we carried out simulations with four common covariate distributions, all with identical first two moments: Normal, $p(z_i) = \mathrm{N}(0,1)$, Rademacher, $p(z_i) = \frac{1}{2}\delta(z_i-1) + \frac{1}{2}\delta(z_i+1)$, Uniform, $z_i \sim \mathrm{U}(-\sqrt{3},\sqrt{3})$ and the Student t-distribution, $z_i \sim \mathrm{t{-}dist}(\nu)/\sqrt{\nu/(\nu-2)}$ (with degrees of freedom $\nu = 5$). The deviations between predictions using Gaussian covariates and the above distributions were indeed small ($< 1\%$ for w and $< 0.5\%$ for v - see Figure 4.3) validating our asymptotic assumption that our theory admits a range of covariates distributions.

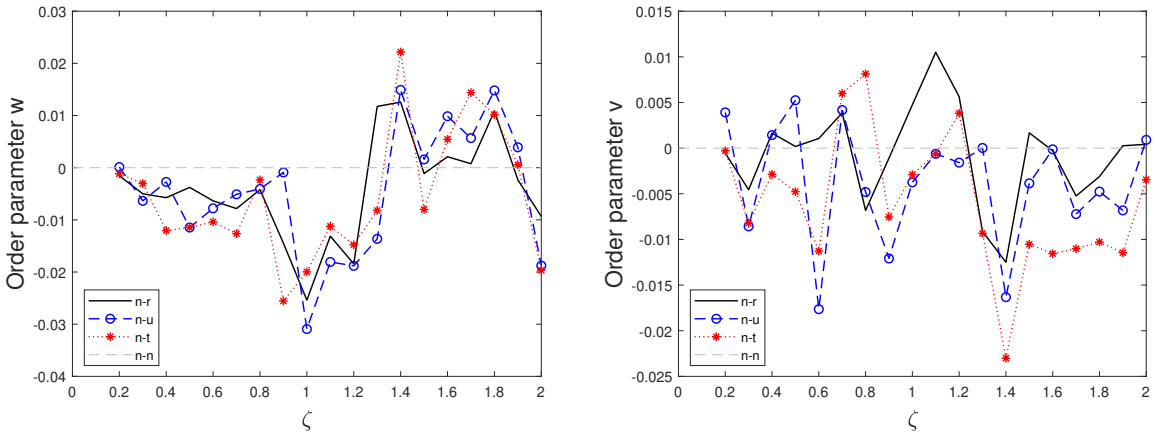

Fig. 4.3 Deviations between simulations using Gaussian covariates and other distributions are shown for $\zeta \in (0,2]$. The four covariate distributions, all with identical first two moments, are Normal (n), $p(z_i) = \mathrm{N}(0,1)$, Rademacher (r), $p(z_i) = \frac{1}{2}\delta(z_i-1) + \frac{1}{2}\delta(z_i+1)$, Uniform (u), $z_i \sim \mathrm{U}(-\sqrt{3},\sqrt{3})$ and the Student t-distribution (t), $z_i \sim \mathrm{t-dist}(\nu)/\sqrt{\nu/(\nu-2)}$ (with degrees of freedom $\nu = 5$). Recall the order parameters $w, v \sim \mathscr{O}(1)$ so the deviations are small.

4.3.2 Uncorrelated covariates

Numerical solution of the RS saddle point equations (4.36a)–(4.36g), with the variational approximation for the base hazard rate, results in data as shown in Figure 4.4. This figure corresponds to $\mathbf{A} = \mathbb{I}$, i.e. uncorrelated and normalized covariates, and $S = 1$. The phase transition at $\zeta = 1$ corresponding to ML regression is for $\eta > 0$ no longer present. As η increases, we find the slope κ (which for the present parameter settings is identical to w) and the variance v of the data cloud decreases.

To test the above predictions, we generated synthetic time-to-event data using zero mean covariate vectors $\mathbf{z}$ with covariance matrix $\mathbf{A}$, and Gaussian random and zero-average association vectors β^0, for different values of N and p. Base hazard rates were chosen to be constant and event times were generated from the Cox proportional hazards model following [10]. From the simulated data we then extracted estimates of the association parameters via penalized Cox regression (using the R package, *glmnet* [56]). Upon solving our RS order parameter equations (4.36a)–(4.36g) for the chosen values of $\zeta = p/N$ and $S^2 = p^{-1}(\beta^0)^2$, we compared against simulations. By construction, there is no model mismatch, since the data are generated from the model assumed in parameter inference. Our theoretical predictions for the slope and variance agree remarkably well with the simulations; see Figure 4.4. In addition, we solve the RS order parameter equations numerically, for a fixed $S = 1$ and $\mathbf{A} = \mathbb{I}$ but varying η, to investigate the effect of regularization (shown in Figure 4.5).

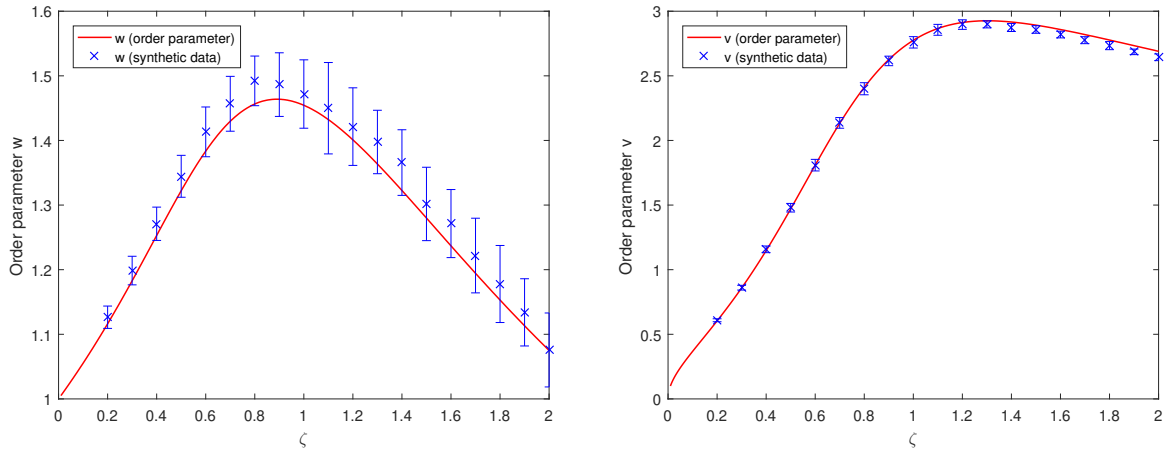

Fig. 4.4 Predicted and measured values of the order parameters w and v (solid lines and markers, respectively), for $\mathbf{A} = \mathbb{I}$, $S = 1$ and $p = 2000$, shown versus $\zeta = p/N \in (0,2]$. Measurements are determined via MAP regression, with regularization parameter $\eta = 0.025$. Simulations are repeated 50 times with independent data sets (generated according to [10], with constant hazard rates), and results shown as averages with error bars indicating one standard deviation. Note that for these settings, slope and the width of the association parameter cloud equal w and v, respectively.

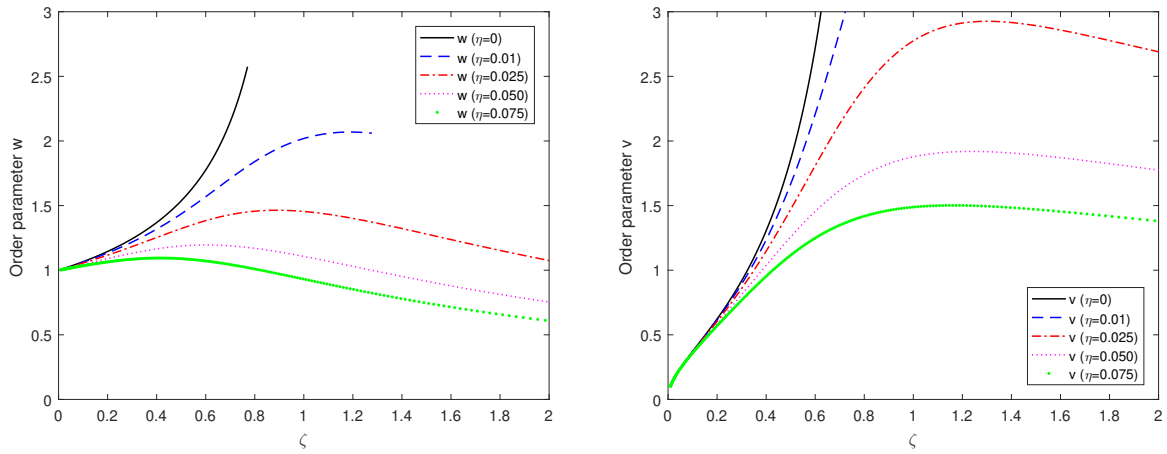

Fig. 4.5 Predicted values of the order parameters w (left) and v (right), shown versus $\zeta = p/N$. They are obtained by solving numerically the RS equations (4.36a)–(4.36g) for $\mathbf{A} = \mathbb{I}$ and $S = 1$, with the variational approximation for $\lambda(t)$, and different choices of the regularization parameter η.

4.3.3 Correlated covariates

Our theory allows us to investigate correlated data by examining the conditions under which expressions of the form $\lim_{p\to\infty} p^{-1}\beta\cdot\mathbf{A}\beta$ are self-averaging. The resulting conditions on the eigenvalue spectrum $\rho(a)$ of $\mathbf{A}$ determine suitable test covariance matrices to use in our simulations. We proceed by considering two non-diagonal covariance matrices $\mathbf{A}$, both with $\lim_{p\to\infty}\langle a\rangle = 1$ and $\lim_{p\to\infty}\langle a^2\rangle = 1+\varepsilon^2$ (hence with spectra of finite width), and $\varepsilon = \mathcal{O}(1)$.

Our first choice was $A_{\mu\nu} = \delta_{\mu\nu} + (1-\delta_{\mu\nu})\varepsilon/\sqrt{p}$, with eigenvalues $1-\varepsilon/\sqrt{p}$ (multiplicity $p-1$) and $1+(p-1)\varepsilon/\sqrt{p}$ (multiplicity 1).

$$A = \begin{pmatrix} 1 & \varepsilon/\sqrt{p} & \dots & \dots & \varepsilon/\sqrt{p} \\ \varepsilon/\sqrt{p} & 1 & \dots & \dots & \varepsilon/\sqrt{p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon/\sqrt{p} & \varepsilon/\sqrt{p} & \dots & \dots & 1 \end{pmatrix} \tag{4.44}$$

A matrix where each covariate is correlated with all others is a rigorous test for our theory but upon working out the spectrum-dependent quantities in the RS equations, we find that for this matrix choice they are independent of ε (see appendix B.9). Hence the order parameters are predicted to be identical to those for data with uncorrelated covariates. Simulations (not shown here) confirm that this is indeed the case, modulo finite size fluctuations.

Our second choice for $\mathbf{A}$ had again $A_{\mu\mu} = 1$ for all μ, but now covariates are correlated in ordered pairs: $A_{\mu,\mu+1} = A_{\mu+1,\mu} = \varepsilon$ for all μ odd, with $A_{\mu\nu} = 0$ for all other $\mu \neq \nu$ (with $0 \leq \varepsilon \leq 1$). This is a block diagonal matrix with $\rho(a) = \frac{1}{2}\delta(a-1-\varepsilon) + \frac{1}{2}\delta(a-1+\varepsilon)$, and the RS order parameters *will* depend on the strength ε of the covariate correlations (see appendix B.9).

$$A = \begin{pmatrix} 1 & \varepsilon & 0 & 0 & \dots & 0 \\ \varepsilon & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \varepsilon & \dots & 0 \\ 0 & 0 & \varepsilon & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & \varepsilon \\ 0 & 0 & 0 & 0 & \varepsilon & 1 \end{pmatrix} \tag{4.45}$$

For both covariance matrix cases, increasing the parameter ε increases the amount of covariate correlation. In Figure 4.6, we show the values of the order parameters v and w, as solved from the RS equations, for $S = 1$, $\eta = 0.025$ and different values of the correlation parameter ε, as functions of ζ. Here we again have $\kappa = w$ and use covariance matrix defined

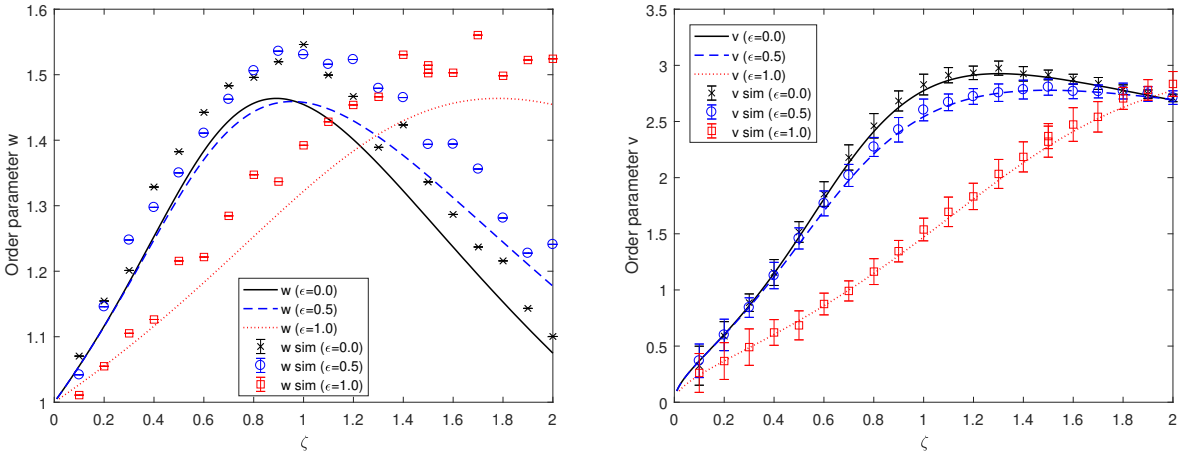

Fig. 4.6 Predicted values of the order parameters w (left) and v (right), shown versus $\zeta = p/N$. They are obtained by solving numerically the RS equations (4.36a)–(4.36g) for $\eta = 0.025$ and $S = 1$, with the variational approximation for $\lambda(t)$. Here the covariates are pairwise correlated according to $A_{\mu,\mu+1} = A_{\mu+1,\mu} = \varepsilon$ for all μ odd, with $A_{\mu\nu} = 0$ for all other $\mu \neq \nu$, with $\varepsilon \in [0,1]$. Note that for these settings, w and v are the slope and the width of the association data cloud. For the left w plot, only mean simulation values are shown since including the errors bars of approximately $\pm 10\%$ led to cluttered plots. Error bars can be displayed clearly for all values of ε on the right v plot. The markers each represent averages over 32 regressions with distinct covariate and association realizations and we fix the value of $Np = 400,000$.

in (4.45). In the same figure we show the results of numerical simulations carried out for $\varepsilon = \{0.0, 0.5, 1.0\}$ and $Np = 400,000$. The error bars of approximately $\pm 10\%$ were not displayed for clarity. The covariates were generated according to: $z_\mu^i = y_{i\mu}$ for μ odd, and $z_\mu^i = \varepsilon y_{i\mu-1} + \sqrt{1-\varepsilon^2}\, y_{i\mu}$, in which all $\{y_{i\mu}\}$ are independent Gaussian random variables, with $\langle y_{i\mu} \rangle = 0$ and $\langle y_{i\mu}^2 \rangle = 1$. This choice generates the above covariate correlations $A_{\mu\nu}$. The markers each represent averages over 32 regressions with distinct covariate and association realizations. The agreement between theory and simulations is seen to be quite satisfactory. We observe that the effect of covariate correlations on the overfitting noise is always a reduction (v decreases with ε).

4.3.4 Alternative to cross-validation

In MAP analyses the regularization parameter η is usually determined by k-fold cross-validation, or via the Generalized Cross Validation (GCV) estimator [36]. A fraction of the data is set aside for this purpose, leaving fewer samples available for inference of model parameters. This has a detrimental effect on inference accuracy. Our present theory, in

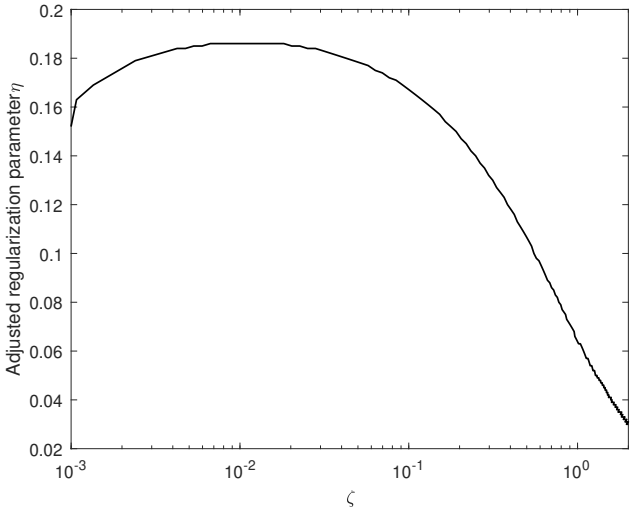

Fig. 4.7 The present theory allows for the analytical identification of the optimally adjusted MAP regularization parameter for Cox regression, by solving the RS order parameter equations (4.36a)–(4.36g) upon demanding unbiased recovery of regression coefficients, $\kappa = 1$, with η as parameter to be solved instead of w. Here we show the result versus $\zeta = p/N$. It is not straightforward to solve the order parameter equations close to $\zeta = 0$, but we know that the curve should tend to the origin for $\zeta = 0$ (where ML inference is asymptotically exact).

contrast, suggests a more data efficient method of estimating the amount of regularization needed for an average dataset, without the need to sacrifice any samples. By fixing the slope parameter to unbiased recovery of the regression coefficients, i.e. $w/\tilde{S} = 1$, and solving the order parameter equations (4.36a)–(4.36g) with η as a parameter to be determined (instead of w), the *optimal* values of η can be estimated without any cross-validation; see Figure 4.7. The optimal values in Figure 4.7 are seen to match those (ζ, η) pairs in Figure 4.4 where $w/\tilde{S} = 1$, as they should. For example, when $\zeta = 1$, the required amount of regularization to compensate for high covariate dimensionality can be read off from Figure 4.7 to be $\eta \approx 0.05$. In interpreting this figure, however, we should note our rescaling of our association parameters, prompted by the observation that $\beta^2 = \mathscr{O}(1)$ is required to avoid non-finite event times for large p. This implied that our $L2$ prior in MAP inference is of the form $p(\beta) \propto \exp(-\eta p \beta^2)$.

The upward sloping region of Figure 4.7, for small ζ, matches our intuition of requiring an increasing amount of regularization for an increasing ζ (up to $\zeta \approx 0.01$). However, as ζ is increased further, we see that optimal regularization now requires a decreasing value of η. To test this less intuitive prediction, we chose four larger values of ζ, read off the required values of η for unbiased inference from Figure 4.7, and calculated the slope of

the association parameter cloud from 100 simulations. These predictions were made with $p = 250$ suggesting our theory is valid for relatively low values of p and N. The results show that the slope of the association parameter cloud is indeed unity, i.e. for the η values proposed by the RS theory, the overfitting-induced inference bias is indeed suppressed as predicted; see the table below:

ζ	required η	corresponding *glmnet* λ	mean slope $\pm$ 1 s.d
0.110	0.165	0.036	1.007 ± 0.028
0.552	0.100	0.110	1.009 ± 0.081
1.055	0.062	0.131	1.013 ± 0.094
2.001	0.031	0.124	0.956 ± 0.139

Fig. 4.8 Validation that the proposed values of ζ and η result in a slope of one indicating perfect regression. Optimal values of η are calculated by solving the relevant order parameter equations (4.36a)–(4.36g) fixing $w = 1$.

Note that Figure 4.7, together with our confirmation in regression simulations that the predicted optimal values of η indeed induce unbiased MAP estimators for regression coefficients (i.e. slopes $\kappa = 1$ in the association parameter clouds), confirm a posteriori the correctness of the chosen scaling with p of our *L2* prior $p(\beta) \propto \exp(-\eta p \beta^2)$. In those situations where the conditions for our theory to apply are not met, other properties may of course affect the optimal value of η. For instance, our simulated data are generated from the Cox model where the ground truth association vector β^0 is not sparse. Equally, the data could be generated from a model with fewer nonzero associations, including choices for which the Central Limit Theorem no longer guarantees that the risk scores $\beta^0 \cdot \mathbf{z}$ have Gaussian statistics.

4.4 Discussion

Failure to correct multivariate ML or MAP regression results for overfitting can lead to serious inference errors. The inferred regression coefficients of the multivariate Cox model are known to be increasingly biased as the ratio of data dimension p to the sample size N increases. For medical time-to-event analysis, where it is possible to obtain (and common to have) large numbers of measurements per patient, such as genomic, epigenetic and imaging data, this bias is quite problematic. It induces false positive associations, which will inevitably turn out to be non-reproducible. This leads to a preventable waste of time and health funds, and frustrates the translation of the significant progress made in recent decades in medical data acquisition into effective data-driven personalized medicine. In this chapter, which builds on the recent study [26], we have built successfully a theory to predict this bias for the multivariate Cox

model in the presence of ridge regularization, when the data dimension scales as $p \sim N$. This paves the way further for effective overfitting corrections in multivariate MAP inference. Alternatively, our analysis allows for a straightforward analytical determination of the optimal regularization needed to correct the overfitting bias, without having to sacrifice valuable training data to cross-validation. In addition to overfitting-induced inference bias, there is a further effect of overfitting on inferred error bars. To determine the statistical significance of inferred regression coefficients, p-values are typically used. These rely on asymptotic results which do not hold in the regime where both p and N are large with $\zeta = p/N \sim \mathcal{O}(1)$ [44, 133], leading to incorrect rejections of the null hypothesis. Our theory shows that the variance of the inferred regression coefficients around the true value is a function of p/N, necessitating an adjustment to traditional test statistics used in p-value calculations.

The aim of our theory is to provide epidemiologists and clinical trials practitioners with a means of analysing data where $p \sim \mathcal{O}(N)$. This paper considers a student-teacher learning problem where we assume the data-generating model is known. A practical overfitting correction protocol for multivariate MAP regression on high-dimensional time-to-event data, based on our present theory, requires knowledge of the values of S (the magnitude $|\beta^0|$ of the true association parameter vector) and of the eigenvalue spectrum of the covariate correlation matrix $\mathbf{A}$. For synthetic data, these are available by assumption. For real data, S can be computed from the inferred regression parameters $\hat{\beta}$, alongside the RS order parameters, using (2.59), from which one infers the relation $v^2+w^2 = \hat{\beta} \cdot \mathbf{A}\hat{\beta}$ (in non-rescaled notation). The value of $\hat{\beta}$ is available in practice as it is the outcome of the regression. We typically only have access to the empirical covariance matrix from which to infer the covariate correlation matrix $\mathbf{A}$. A possible solution for this problem is to use the link between the empirical and population level eigenvalue distributions in the Marčenko-Pastur equation [94]. The population spectrum can be estimated from its empirical counterpart in [42], by applying convex optimisation to the inverted Marčenko-Pastur equation. This method is an improvement on naively using the sample eigenvalue spectrum as an estimator of its population counterpart when $p \sim N$ (see Figures 1.2, 1.3).

There are many directions for extension of the present line of research. For instance, time-to-event data, whether from observational studies or clinical trials, are typically censored. Censoring may reflect the impact of competing risks, patients withdrawing from studies, or finite study durations. The incorporation of censoring into our theory is an obvious next research target, together with investigation of regimes where the risk score are no longer Gaussian distributed. Finally, the overfitting measure in (4.18) is quite general, and can be applied to many other survival analysis models [85]. Equally, the theory developed in this

paper is directly applicable to time-to-event studies outside medical data such as credit risk analysis.

Part II

Integrable Bayesian Inference in High Dimensions

Chapter 5

Introduction to Bayesian classification

In the previous chapters, we were able to switch between frequentist and Bayesian viewpoints via the regularization parameter η and we approximated from full Bayesian inference to MAP inference. In this section, based on the peer-reviewed article [124], we explicitly pursue the fully Bayesian approach to make inferences about a dataset when $p \sim \mathscr{O}(N)$. So far, the key object has been the likelihood function $p(\mathscr{D}|\vartheta)$ where $\mathscr{D}$ represents the data, typically of the form $\{(\mathbf{x_i}, y_i)\}_{i=1}^{N}$, and ϑ, the vector of model parameters. To infer these model parameters ϑ, the function is maximized [49] with respect to those parameters to find their *true values*.

$$\hat{\vartheta}_{ML} = \underset{\vartheta}{\operatorname{argmax}}\, p(\mathscr{D}|\vartheta) \tag{5.1}$$

In the Bayesian approach, it is acknowledged that there are a range of parameter values consistent with the data [75]. This uncertainty is incorporated through a prior distribution and Bayes Theorem enables a posterior probability distribution $p(\vartheta|\mathscr{D}, \mathscr{H})$ to be calculated.

$$p(\vartheta|\mathscr{D}, \mathscr{H}) = \frac{p(\mathscr{D}|\vartheta, \mathscr{H})\; p(\vartheta|\mathscr{H})}{p(\mathscr{D}|\mathscr{H})} \tag{5.2}$$

where $\mathscr{H}$ are the parameters of the prior distribution (so-called hyperparameters). The likelihood $p(\mathscr{D}|\vartheta, \mathscr{H})$ is the probability of the observed data conditioned on the model parameters (hence a known underlying model) and the hyperparameters. In this thesis, we only consider the case where the likelihood matches the true data-generating model i.e. the model is correctly specified. Prior information is incorporated into the model via $p(\vartheta|\mathscr{H})$ and setting this to a uniform distribution recovers the maximum likelihood case. Finally the evidence term in the denominator, which normalizes the posterior probability distribution, is defined as

$$p(\mathscr{D}|\mathscr{H}) = \int \mathrm{d}\vartheta\; p(\mathscr{D}|\vartheta, \mathscr{H})\; p(\vartheta|\mathscr{H}) \tag{5.3}$$

The prior and posterior distributions are *conjugate* when they belong to the same family of distributions. Examples of conjugacy include the binomial/beta distributions for discrete random variables and multivariate normal/normal-Wishart for continuous random variables. The use of conjugate priors allows for convenient updating of posterior parameters leading to the analytically tractable calculations of early work [58, 82]. The accusation of subjectivity led to the study of maximum entropy priors [74], the transformation invariant Jeffrey's prior [76] and reference priors [12] which maximize the Kullback-Leibler divergence between the prior and the posterior. These proved difficult to work with, for all but the most simple distributions. When analytically convenient priors are replaced with alternatives, symbolical integration is often not possible. We must resort to methods of approximation which can be either deterministic (e.g. variational inference) or stochastic (sampling methods such Markov Chain Monte Carlo) in nature [13].

The main stochastic method used is Markov Chain Monte Carlo (MCMC) or variants such as Gibbs sampling. The aim is to construct a Markov chain whose equilibrium distribution is the distribution of interest. In the Bayesian setting, this target is typically the posterior distribution. The combination of MCMC methods and an explosion in computational power led to the widespread adoption of numerical methods in Bayesian analysis. In turn, this broadened the range of feasible prior choices from conjugate to non-conjugate distributions. The ubiquity of sophisticated computational methods has meant analytical approaches have somewhat fallen out of favour.

5.1 Classification

Bayesian methods can naturally be applied to the problem of classification which maps data samples $\mathbf{x} \in \mathbb{R}^p$ to discrete classes $y \in \{1, \ldots C\}$, by inferring the underlying statistical regularities from a given training set $\mathscr{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ of N independent and identically distributed (i.i.d.) samples and the corresponding classes. In other words, we seek a function from $\mathbb{R}^p \to \{1, \ldots C\}$ which will allow us to allocate a new data point $\mathbf{x}_0$ to a single class y_0. Since the samples have class labels, classification is a supervised learning method.

5.1.1 Discriminative or Generative?

Following the argument of [13], the conditional probability of a class label given a data sample can be written to neatly highlight the difference between discriminative and generative classifiers. Assume a dataset with two discrete classes labelled C_1 and C_2. According to

Bayes Theorem, the required conditional probability can be written as

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_1)p(C_1)+p(\mathbf{x}|C_2)p(C_2)} = \frac{1}{1+\frac{p(\mathbf{x}|C_2)p(C_2)}{p(\mathbf{x}|C_1)p(C_1)}} = \frac{1}{1+e^{-g_{12}}} = \sigma(g_{12}(\mathbf{x})) \tag{5.4}$$

where $g_{12}(\mathbf{x}) = \ln\left\{\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}\right\}$ and $\sigma(g_{12}(\mathbf{x}))$ is a sigmoid function. Generative classifiers assume a specific form for the class-conditional probabilities $p(\mathbf{x}|C_i)$ and the prior probabilities $p(C_i)$. The name arises from the ability to generate new samples from the joint distribution $p(\mathbf{x},C) = p(\mathbf{x}|C)p(C)$. On the other hand, if we infer the parameters of $p(C_1|\mathbf{x})$ directly without appealing to the class-conditional probabilities, the model is called a *discriminative classifier*. A common example is the logistic regression model analyzed in Part I (see (3.1a)). Other forms of discriminative classifier include Support Vector Machines and neural networks. We now provide further detail on generative models since it is the focus of the following chapter.

5.1.2 Generative methods

Generative classification entails defining a suitable parametrization $p(\mathbf{x},y|\vartheta)$ of the multivariate distribution from which the samples in $\mathscr{D}$ were drawn. If the assumed form of the class-conditional probability density functions are multivariate Gaussian, $p(\mathbf{x}|C) = (2\pi)^{-\frac{p}{2}}|\Sigma_C|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\mu_C)^T\Sigma_C^{-1}(\mathbf{x}-\mu_C)}$ where μ_C, Σ_C represent the unknown class mean and covariance matrix, the explicit expression for the discriminant function $g_{12}(\mathbf{x})$ from (5.4) becomes

$$\begin{aligned} g_{12}(\mathbf{x}) = &-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)+\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2) \\ &+\frac{1}{2}\ln|\Sigma_2|-\frac{1}{2}\ln|\Sigma_1|+\ln\frac{p(C_1)}{p(C_2)} \end{aligned} \tag{5.5}$$

where μ_y and Σ_y are the mean vector and covariance matrix specific to class y. A crucial role is played by these class-specific sample covariance matrices, that capture the correlations between the components of $\mathbf{x}$. These are inferred in the frequentist approach or integrated over in the full Bayesian treatment. The class-specific covariance matrices can be used to further categorize the generative approach:

- If the two covariance matrices are identical, $\Sigma_1 = \Sigma_2$, the quadratic terms in $\mathbf{x}$ from (5.5) cancel resulting in a *linear discriminant analysis (LDA)*. This corresponds to a hyperplane where the posterior probabilities of each class are equal and the priors play the role of a constant bias term.

- If $\Sigma_1 \neq \Sigma_2$, the quadratic terms in $\mathbf{x}$ remain resulting in a *quadratic discriminant analysis (QDA)*.

The earliest approaches to generative Bayesian classifiers with a multivariate Gaussian sampling distribution and the analytically tractable conjugate prior were [82, 58]. In this case, the Wishart distribution is the conjugate prior. Assuming a p-dimensional vector $\mathbf{x_i} \sim \mathrm{N}_p(0,\mathbf{S})$, the $p \times p$ symmetric matrix, $\Lambda = r^{-1}\sum_{i=1}^{r} \mathbf{x_i}\mathbf{x_i}^T$ has a Wishart distribution with r degrees of freedom and seed (or scale) matrix, $\mathbf{S}$ (see A.2 for further details). The hyperparameters to specify are the degrees of freedom, r, and the so-called seed matrix, $\mathbf{S}$. The term Quadratic Bayes was introduced by [16] for a seed matrix of the form $k\mathbb{I}$ where $k \in \mathbb{R}^+$. The hyperparameter, k, is determined by cross-validation. A simpler method is to use a "plug-in" estimator e.g. $\mathbf{S} = p^{-1}Tr(\hat{\Sigma}_{ML})\mathbb{I}$ which is still defined for $N < p$ [127, 128]. For this form, there are no additional parameters to estimate. A distribution based prior method is used in [127, 128] rather than the typical parametrized prior.

Mean vectors and covariance matrices can be estimated from the sample data using Maximum Likelihood (ML) estimators. However, as we have seen in the introduction, when the number of observations in class z, N_z, is less than the dimension, p, the estimation problem is ill-posed for covariance matrix estimation in QDA (since Σ_1 and Σ_2 need to be estimated). When the total number of observations, $N = \sum_{z=1}^{C} N_z$, is less than p, estimation is an ill-posed problem for LDA. The regularization method of [54] addresses this problem by introducing regularization parameters which subsequently need to be estimated from the training data.

5.1.3 Hyperparameter estimation

Bayesian inference removes the need to estimate the model parameters ϑ by integrating over them. This necessitates a choice of prior which is itself a probability distribution with parameters $\mathscr{H}$ (so-called hyperparameters). The resulting predictive probability, $p(y_0|\mathbf{x}_0, \mathscr{D}, \mathscr{H})$, which quantifies the probability of class y_0 given data sample $\mathbf{x}_0$, is a function of the data $\mathscr{D}$ and the hyperparameters $\mathscr{H}$ (6.1). The latter can be estimated from the data (empirical Bayes [41, 65]) or be distributed via a hyperprior (hierarchical Bayesian modelling). These hyperpriors are often not parametrized or are relatively uninformative[1] in order to break the inference hierarchy. The role of Bayesian priors and their associated hyperparameters becomes increasingly important as the data dimensionality increases.

Clear protocols for estimating hyperparameters are important for a number of reasons:

1. Accurate estimation improves the performance of the classification algorithm.

[1] For example a uniform distribution with a large but fixed variance, $\mathbf{U}(-1000, 1000)$

2. New classification methods are frequently compared to existing ones in the scientific literature but often little attention is paid to the important step of hyperparameter estimation. Methods can be different between classifiers or not fully described at all leading to an unfair comparisons.

3. The lack of explicit hyperparameter estimation methodology in many academic papers prevents reproducible results even when the exact classifier and data is known.

Our methodology is derived via evidence maximization and described in Section 6.3.

5.2 Applications to medical data

The use of classification algorithms for diagnostic and prognostic medical applications is now commonplace. Often a combination of feature selection and a classifier are used to discriminate between biological samples e.g. logistic/probit regression [2, 47, 143], Naive Bayes[2] [83], Empirical Bayes [90] and fully Bayesian hierarchical methods [73]. Rationales for selecting a subset of genes include reducing overfitting, increasing classification speed and providing potential for further experimental study. Methods range from standard approaches such as principle component analysis [111] to bespoke methods which take into account biological knowledge [21, 24]. Combinatorial approaches are used [147, 148] to account for interactions between large numbers of genes.

If the number of samples N is large compared to the data dimensionality p, computing point estimates of the unknown parameters ϑ by maximum likelihood (ML) is accurate and usually sufficient. On the other hand, if the ratio p/N is different from zero, point estimation based methods are prone to overfitting. This is the "curse of dimensionality". Unfortunately, the regime of finite p/N is increasingly relevant for modern medical applications, where clinical datasets often report on relatively few patients but contain many measurements per patient[3].

In the following chapter, we avoid the use of feature selection by developing a Bayesian method which considers all possible variable combinations. This is possible by judicious model selection allowing a full analytical treatment.

[2] Naive Bayes assumes that each covariate is independent of the others given the class label, C_z.

[3] This is the case for rare diseases, or when obtaining tissue material is nontrivial or expensive, but measuring extensive numbers of features in such material (e.g. gene expression data) is relatively simple and cheap.

Chapter 6

Accurate Bayesian Data Classification without Hyperparameter Cross-validation

6.1 Introduction

From the general Bayesian framework (5.2)(5.3) and assumptions for the class-conditional probability densities, we can proceed with our aim of calculating the posterior and predictive probabilities either numerically or analytically. Despite all the advantages of the former, approximating the posterior distribution numerically becomes problematic when considering a large space of models. In addition, iterations stuck in local modes often require heuristic solutions. Before the wide-spread usage of MCMC, Bayesian analysis was restricted to simple models and their conjugate priors. By resurrecting and further elaborating methods used at this time, we attempt to make progress by widening the family of analytical priors. We motivate this focus on the prior by noting that as the data dimensionality increases, so does the role of the Bayesian prior. The Wishart distribution is the canonical prior for the covariance matrices. Analytically tractable choices for the class means are the conjugate [58, 82] or the non-informative priors [16, 127]. Our approach allows us to derive closed form expressions for the predictive probabilities of two special model instances with all integrals whose dimensions scale with p being solved analytically. We work under the premise that this may still be preferable to sampling from a very high-dimensional posterior distribution using MCMC or Gibbs sampling.

Any sensible generative model for classifying vectors in $\mathbb{R}^p$ will have at least $\mathcal{O}(p)$ parameters. The fundamental cause of overfitting is the fact that in high-dimensional spaces,

where p/N is finite even if N is large, the posterior parameter distribution $p(\vartheta|\mathscr{D})$ (in a Bayesian sense) will be extremely sparse. Replacing this posterior by a delta-peak, which is what point estimation implies, is always a very poor approximation, irrespective of which protocol is used for estimating the location of this peak. It follows that by retaining the full posterior distribution and doing all integrations over model parameters *analytically*, one should reduce overfitting effects, potentially allowing for high-dimensional datasets to be classified reliably. The need to evaluate all parameter integrals analytically limits us in practice to parametric generative models with class-specific multivariate Gaussian distributions. Here the model parameters to be integrated over are the class means in $\mathbb{R}^p$ and class-specific $p \times p$ covariance matrices, and with carefully chosen priors one can indeed obtain analytical results.

Having completed the relevant integrals, we are left to estimate the hyperparameters, whose dimensionality is normally small, and independent of p. The most commonly used route for hyperparameter estimation appears to be cross-validation which requires re-training one's model k times for k-fold cross-validation; for leave-one-out cross-validation, the model will need to be re-trained N times. Fortunately, a by-product of evaluating our Bayesian integrals analytically is obtaining a system of equations for the class-specific hyperparameters. This is achieved through evidence maximization [91]. Solving these low-dimensional equations avoids the need for resorting to computationally expensive cross-validation methods, generalized cross-validation estimators [36] or subjective hyperparameter choices. The behaviour of the optimal hyperparameters is explored in the high-dimensional data regime. The classification accuracy of the resulting generalized model is competitive with state-of-the art Bayesian discriminant analysis methods, but without the usual computational burden of cross-validation.

In Section 6.2 we define our generative Bayesian classifier and derive the relevant integrals. Special analytically solvable cases of these integrals, leading to two models (A and B), are described in Section 6.3 along with the evidence maximization estimation of hyperparameters. Closed form expressions for the predictive probabilities corresponding to these two models are obtained in Section 6.3.4. We then examine the behaviour of the hyperparameters in Section 6.4.1, and carry out comparative classification performance tests on synthetic and real datasets in Section 6.5. We conclude this chapter with a discussion of the main results. It is worth noting the statistical physics methods of Part I required the limit $p, N \to \infty$ to be taken whereas the Bayesian methods of Part II do not have any asymptotic assumptions.

6.2 Definitions

6.2.1 Model and objectives

We have data $\mathscr{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ where $\mathbf{x_i} \in \mathbb{R}^p$ is a vector of covariates, and $y_i \in \{1, \ldots, C\}$ a discrete outcome label. We seek to predict the outcome y_0 associated with a *new* covariate vector $\mathbf{x}_0$, given the data $\mathscr{D}$. The required predictive probability to compute is therefore

$$\begin{aligned} p(y_0|\mathbf{x}_0, \mathscr{D}) &= \frac{p(y_0, \mathbf{x}_0|\mathscr{D})}{\sum_{y=1}^{C} p(y, \mathbf{x}_0|\mathscr{D})} = \frac{p(y_0, \mathbf{x}_0|\mathbf{x}_1, \ldots, \mathbf{x}_N; y_1, \ldots, y_N)}{\sum_{y=1}^{C} p(y, \mathbf{x}_0|\mathbf{x}_1, \ldots, \mathbf{x}_N; y_1, \ldots, y_N)} \\ &= \frac{p(\mathbf{x}_0, \ldots, \mathbf{x}_N; y_0, \ldots, y_N)}{\sum_{y=1}^{C} p(\mathbf{x}_0, \ldots, \mathbf{x}_N; y, y_1, \ldots, y_N)} \end{aligned} \tag{6.1}$$

where the denominator in the first equality is simply the marginal i.e. $\sum_{y=1}^{C} p(y, \mathbf{x}_0|\mathscr{D}) = p(\mathbf{x}_0|\mathscr{D}) = p(\mathbf{x}_0)$ as $\mathbf{x}_0$ is independent of $\mathscr{D}$. Since we are creating a generative classifier, we start from an expression for the joint distribution $p(\mathbf{x}_0, \ldots, \mathbf{x}_N; y_0, \ldots, y_N)$ and assume that all pairs $(\mathbf{x}_i, y_i)$ are drawn independently from a parametrized distribution $p(\mathbf{x}, y|\vartheta)$ whose parameters ϑ we don't know. Therefore the joint distribution of $\{(\mathbf{x}_i, y_i)\}_{i=0}^{N}$ factorizes

$$p(\mathbf{x}_0, \ldots, \mathbf{x}_N; y_0, \ldots, y_N) = \int_{\vartheta \in \Theta} \mathrm{d}\vartheta \; p(\vartheta) \prod_{i=0}^{N} p(\mathbf{x}_i, y_i|\vartheta) \tag{6.2}$$

where $p(\vartheta)$ is the prior probability distribution of the model parameters. All ϑ integrals are over Θ until a specific model is specified. It now follows that

$$p(y_0|\mathbf{x}_0, \mathscr{D}) = \frac{\int \mathrm{d}\vartheta \; p(\vartheta) \prod_{i=0}^{N} p(\mathbf{x}_i, y_i|\vartheta)}{\sum_{y=1}^{C} \int \mathrm{d}\vartheta \; p(\vartheta) p(\mathbf{x}_0, y|\vartheta) \prod_{i=1}^{N} p(\mathbf{x}_i, y_i|\vartheta)} \tag{6.3}$$

We regard all model parameters with dimensionality that scales with the covariate dimension p as *micro-parameters*, over which we need to integrate (in the sense of ϑ above). Parameters with p-independent dimensionality are regarded as *hyperparameters*. The hyperparameter values will be called a 'model' $\mathscr{H}$. Our equations will now be conditioned on the label $\mathscr{H}$:

$$p(y_0|\mathbf{x}_0, \mathscr{D}, \mathscr{H}) = \frac{\int \mathrm{d}\vartheta \; p(\vartheta|\mathscr{H}) \prod_{i=0}^{N} p(\mathbf{x}_i, y_i|\vartheta, \mathscr{H})}{\sum_{y=1}^{C} \int \mathrm{d}\vartheta \; p(\vartheta|\mathscr{H}) p(\mathbf{x}_0, y|\vartheta, \mathscr{H}) \prod_{i=1}^{N} p(\mathbf{x}_i, y_i|\vartheta, \mathscr{H})} \tag{6.4}$$

where y_0 is only dependent on $\mathscr{H}$ through ϑ.

An aside on evidence. The evidence term introduced in (5.3) can be expressed as

$$p(\mathscr{D}|\mathscr{H}) = p(\mathbf{x}_1,\ldots,\mathbf{x}_N;y_1,\ldots,y_N|\mathscr{H}) = \int \mathrm{d}\vartheta\ p(\vartheta|\mathscr{H}) \prod_{i=1}^{N} p(\mathbf{x}_i,y_i|\vartheta,\mathscr{H}) \tag{6.5}$$

We do not include the new sample $(\mathbf{x}_0, y_0)$ to be classified since it is not part of the data $\mathscr{D}$. Using Bayes Theorem, we can find the probability of $\mathscr{H}$ conditioned on $\mathscr{D}$

$$p(\mathscr{H}|\mathscr{D}) \ = \frac{p(\mathscr{D}|\mathscr{H})p(\mathscr{H})}{\sum_{\mathscr{H}'} p(\mathscr{D}|\mathscr{H}')p(\mathscr{H}')} \tag{6.6}$$

where $p(\mathscr{H})$ is a prior on the hyperparameters – a hyperprior and the summation in the denominator is over all possible discrete hyperparameter values. If $\mathscr{H}$ is a continuous variable, we replace the summation with an integral. The Bayes-optimal hyperparameters $\mathscr{H}$ refer to the situation where the true model is known but the value of the parameters are to be inferred. These maximize the evidence, i.e.

$$\begin{aligned} \hat{\mathscr{H}} = \mathrm{argmax}_{\mathscr{H}}\, p(\mathscr{H}|\mathscr{D}) \ &= \ \mathrm{argmax}_{\mathscr{H}} \log\left\{ \frac{p(\mathscr{D}|\mathscr{H})p(\mathscr{H})}{\sum_{\mathscr{H}'} p(\mathscr{D}|\mathscr{H}')p(\mathscr{H}')} \right\} \\ &= \mathrm{argmax}_{\mathscr{H}} \left\{ \log \int \mathrm{d}\vartheta\ p(\vartheta|\mathscr{H}) \prod_{i=1}^{N} p(\mathbf{x}_i,y_i|\vartheta,\mathscr{H}) + \log p(\mathscr{H}) \right\} \end{aligned} \tag{6.7}$$

We neglect the denominator in the third equality since it is independent on $\mathscr{H}$. There are two possible assumptions for $p(\mathscr{H})$ in the final term of (6.7): a point estimate or a non-parametric probability distribution to break the inference hierarchy (typically a uniform distribution with large but fixed domain). The former is called Empirical Bayes [41, 65]. We choose the latter approach by maximizing the evidence term assuming a flat hyperprior $p(\mathscr{H})$ (details in Section 6.3). Since the $\log p(\mathscr{H})$ term no longer depends on $\mathscr{H}$, it is not relevant in the maximization equation (6.7).

Returning to the predictive probability (6.4), we must specify a parametrization $p(\mathbf{x},y|\vartheta)$ of the joint statistics of covariates $\mathbf{x}$ and class labels y in the population from which our samples are drawn. This choice is constrained by our desire to do all integrations over ϑ analytically, to avoid approximations and overfitting problems caused by point-estimation. One is then naturally led to class-specific Gaussian covariate distributions:

$$p(\mathbf{x},y|\vartheta) = p(y|\vartheta)p(\mathbf{x}|y,\vartheta), \qquad p(\mathbf{x}|y,\vartheta) = \frac{\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu_y)\cdot\Lambda_y(\mathbf{x}-\mu_y)}}{\sqrt{(2\pi)^p/\mathrm{Det}\Lambda_y}} \tag{6.8}$$

Thus the parameters to be integrated over are $\vartheta = \{\mu_y, \Lambda_y,\ y = 1,\ldots,C\}$, i.e. the class-specific means and precision matrices (inverse of the covariance matrices).

The uncertainty in plausible parameter values consistent with the data is incorporated in the prior distribution. However, in this work, we assume the true form of the sampling distribution is known.

6.2.2 Integrals to be computed

In both the inference formula (6.4) and the expression for $\hat{\mathscr{H}}$ (6.7), the relevant integral to be evaluated analytically is

$$\Omega(\mathscr{H}, N, \mathscr{D}) \equiv -\log \int \mathrm{d}\vartheta\ p(\vartheta|\mathscr{H}) \prod_{i=1}^{N} p(\mathbf{x}_i, y_i|\vartheta, \mathscr{H}) \tag{6.9}$$

In the case where we require $\Omega(\mathscr{H}, N+1, \mathscr{D})$, when evaluating the numerator and the denominator of (6.4), we simply replace $\prod_{i=1}^{N}$ by $\prod_{i=0}^{N}$, so that

$$p(y_0|\mathbf{x}_0, \mathscr{D}) = \frac{\mathrm{e}^{-\Omega(\mathscr{H}, N+1, \mathscr{D})}}{\sum_{z=1}^{C} \mathrm{e}^{-\Omega(\mathscr{H}, N+1, \mathscr{D})}|_{y_0=z}} \qquad \hat{\mathscr{H}} = \mathrm{argmin}_{\mathscr{H}} \Omega(\mathscr{H}, N, \mathscr{D}) \tag{6.10}$$

To be clear, there are N samples in the training data so hyperparameter estimation involves $\Omega(\mathscr{H}, N, \mathscr{D})$. The likelihood term used in the calculation of the predictive probability (numerator in (6.4)) requires the additional $(\mathbf{x}_0, y_0)$ pair and hence $N+1$ samples leading to $\Omega(\mathscr{H}, N+1, \mathscr{D})$. Working out $\Omega(\mathscr{H}, N, \mathscr{D})$ for the parametrization (6.8) gives:

$$\begin{aligned}\Omega(\mathscr{H}, N, \mathscr{D}) = \frac{1}{2}Np\log(2\pi) - \sum_{i=1}^{N} \log p_{y_i} \\ - \log \int \Big[\prod_{z=1}^{C} \mathrm{d}\mu_z \mathrm{d}\Lambda_z\ p_z(\mu_z, \Lambda_z)\Big] \Big[\prod_{i=1}^{N} (\mathrm{Det}\Lambda_{y_i})^{\frac{1}{2}}\Big] \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{N} (\mathbf{x}_i - \mu_{y_i}) \cdot \Lambda_{y_i} (\mathbf{x}_i - \mu_{y_i})}\end{aligned} \tag{6.11}$$

where $p(y_i) = p_{y_i}$ is the prior probability of a sample belonging to class y_i. For generative models, this is typically equal to the empirical proportion of samples in that specific class. This is shown explicitly in Section 6.3. To simplify this expression we define the data-dependent index sets $I_z = \{i|\ y_i = z\}$, each of size $N_z = |I_z| = \sum_{i=1}^{N} \delta_{z,y_i}$. We also introduce empirical covariate averages and correlations, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$:

$$\hat{x}^z_\mu = \frac{1}{N_z} \sum_{i \in I_z} x_{i\mu}, \qquad \hat{C}^z_{\mu\nu} = \frac{1}{N_z} \sum_{i \in I_z} (x_{i\mu} - \hat{x}^z_\mu)(x_{i\nu} - \hat{x}^z_\nu) \tag{6.12}$$

Upon defining the vector $\hat{\mathbf{x}}_z = (\hat{x}^z_1,\ldots,\hat{x}^z_p)$, and the $p\times p$ real symmetric matrix $\hat{\mathbf{C}}_z = \{\hat{C}^z_{\mu\nu}\}$, we can then write the relevant integrals after some simple rearrangements in the form

$$\begin{aligned}
\Omega(\mathscr{H},N,\mathscr{D}) &= \frac{1}{2}Np\log(2\pi) - \sum_{z=1}^{C} N_z \log p_z \\
&- \log\int\Big[\prod_{z=1}^{C} \mathrm{d}\mu_z \mathrm{d}\Lambda_z\; p_z(\mu_z,\Lambda_z)(\mathrm{Det}\Lambda_z)^{\frac{N_z}{2}} \mathrm{e}^{-\frac{1}{2}N_z\mu_z\cdot\Lambda_z\mu_z}\Big] \\
&\times\; \mathrm{e}^{\sum_{z=1}^{C}\mu_z\cdot\Lambda_z\sum_{i\in I_z}\mathbf{x}_i - \frac{1}{2}\sum_{z=1}^{C}\sum_{i\in I_z}\mathbf{x}_i\cdot\Lambda_z\mathbf{x}_i} \\
&= \frac{1}{2}Np\log(2\pi) - \sum_{z=1}^{C} N_z\log p_z \\
&- \sum_{z=1}^{C}\log\int \mathrm{d}\mu_z\mathrm{d}\Lambda_z\; p_z(\mu_z+\hat{\mathbf{x}}_z,\Lambda_z)(\mathrm{Det}\Lambda_z)^{\frac{1}{2}N_z}\mathrm{e}^{-\frac{1}{2}N_z\mu_z\cdot\Lambda_z\mu_z - \frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda_z)}
\end{aligned} \tag{6.13}$$

since the normal distribution satisfies the convolution for random variable X_1+X_2

$$X_1\sim \mathrm{N}(\mu_1,\sigma_1^2),\;\; X_2\sim\mathrm{N}(\mu_2,\sigma_2^2) \;\;\Rightarrow\;\; \mathrm{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2) \tag{6.14}$$

and we have used the compact form in expectation of choosing a prior distribution

$$\begin{aligned}
\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda) &= \sum_{\mu=1}^{p}(\hat{\mathbf{C}}_z\Lambda)_{\mu\mu} = \sum_{\mu,\rho=1}^{p}\hat{\mathbf{C}}^z_{\mu\rho}\Lambda_{\rho\mu} = \sum_{\mu\rho}\frac{1}{N_z}\sum_{i\in I_z}(x_{i\mu}-\hat{x}^z_\mu)(x_{i\rho}-\hat{x}^z_\rho)\Lambda_{\rho\mu} \\
&= \frac{1}{N_z}\sum_{\mu\rho}\hat{x}^z_\mu\Lambda_{\rho\mu}\hat{x}^z_\rho + \frac{1}{N_z}\sum_{\mu\rho}\sum_{i\in I_z}\Big\{x_{i\mu}\Lambda_{\rho\mu}x_{i\rho} - x_{i\mu}\Lambda_{\rho\mu}\hat{x}^z_\rho - x_{i\rho}\Lambda_{\rho\mu}\hat{x}^z_\mu\Big\} \\
&= \frac{1}{N_z}\hat{\mathbf{x}}^z\Lambda\hat{\mathbf{x}}^z + \frac{1}{N_z}\sum_{i\in I_z}\Big\{\mathbf{x}_i\Lambda\mathbf{x}_i - 2\hat{\mathbf{x}}^z\Lambda\mathbf{x}_i\Big\}
\end{aligned} \tag{6.15}$$

To proceed it is essential that we compute $\Omega(\mathscr{H},N,\mathscr{D})$ analytically, for arbitrary $\hat{\mathbf{x}}\in\mathbb{R}^p$ and arbitrary positive definite symmetric matrices $\hat{\mathbf{C}}$. This will constrain the choice of our priors $p_z(\mu,\Lambda)$ for the covariate averages and correlations in outcome class z. All required integrals are of the following form, with Λ limited to the subset Ξ^p of symmetric positive definite matrices:

$$\Psi_z(\mathscr{H},N,\mathscr{D}) = \int_{\mathbb{R}^p}\mathrm{d}\mu\int_{\Xi^p}\mathrm{d}\Lambda\; p_z(\mu+\hat{\mathbf{x}}_z|\Lambda)p_z(\Lambda)(\mathrm{Det}\Lambda)^{\frac{1}{2}N_z}\mathrm{e}^{-\frac{1}{2}N_z\mu\cdot\Lambda\mu-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda)} \tag{6.16}$$

We will drop the indications of the sets over which the integrals are done, when these are clear from the context. The tricky integral is that over the inverse covariance matrices Λ. The choice in [122] corresponded to $p_z(\mu,\Lambda)\propto \mathrm{e}^{-\frac{1}{2}\mu^2/\beta_z^2}\delta[\Lambda-\mathbb{I}/\alpha_z^2]$, which implied assuming

uncorrelated covariates within each class. Here we want to allow for arbitrary class-specific covariate correlations.

6.2.3 Priors for class-specific means and covariance matrices

The integrals over μ and Λ can be evaluated in either order. We start with the integral over μ. In contrast to most studies, we replace the conjugate prior for the unknown mean vector by a multivariate Gaussian with as yet arbitrary class-specific precision matrices $\mathbf{A}_z$. This should allow us to cover a larger parameter space than the conjugate prior (which has ${\Lambda_z}^{-1}$ as its covariance matrix):

$$p_z(\mu|\mathbf{A}_z) = (2\pi)^{-\frac{p}{2}}\sqrt{\mathrm{Det}\mathbf{A}_z}\,\mathrm{e}^{-\frac{1}{2}\mu\cdot\mathbf{A}_z\mu} \tag{6.17}$$

Insertion into (6.16) and using the moment generating function for Gaussian random variables, $\mathbb{E}(e^{\mathbf{s}.\mathbf{x}}) = e^{\frac{1}{2}\mathbf{s}.\Sigma\mathbf{s}}$ with $\mathbf{s} = -\mathbf{A}_z\hat{\mathbf{x}}_z$ and $\Sigma = (N_z\Lambda+\mathbf{A}_z)$ (see Appendix A.1) to complete the μ integral gives

$$\begin{aligned}
\Psi_z &= (2\pi)^{-\frac{p}{2}}\int \mathrm{d}\Lambda\; p_z(\Lambda)\mathrm{e}^{-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda)-\frac{1}{2}\hat{\mathbf{x}}_z\cdot\mathbf{A}_z\hat{\mathbf{x}}_z}\left[\mathrm{Det}(\Lambda^{N_z})\mathrm{Det}\mathbf{A}_z\right]^{\frac{1}{2}} \\
&\quad\times\int \mathrm{d}\mu\; \mathrm{e}^{-\frac{1}{2}\mu\cdot(N_z\Lambda+\mathbf{A}_z)\mu-\mu\cdot\mathbf{A}_z\hat{\mathbf{x}}_z} \\
&= \int \mathrm{d}\Lambda\; p_z(\Lambda)\mathrm{e}^{-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda)}\left[\frac{\mathrm{Det}(\Lambda^{N_z})\mathrm{Det}\mathbf{A}_z}{\mathrm{Det}(N_z\Lambda+\mathbf{A}_z)}\right]^{\frac{1}{2}}\mathrm{e}^{\frac{1}{2}\hat{\mathbf{x}}\cdot\mathbf{A}_z(N_z\Lambda+\mathbf{A}_z)^{-1}\mathbf{A}_z\hat{\mathbf{x}}_z-\frac{1}{2}\hat{\mathbf{x}}_z\cdot\mathbf{A}_z\hat{\mathbf{x}}_z} \\
&= \int \mathrm{d}\Lambda\; p_z(\Lambda)\mathrm{e}^{-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda)}\left[\mathrm{Det}(N_z\Lambda^{1-N_z}\mathbf{A}_z^{-1}+\Lambda^{-N_z})\right]^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\hat{\mathbf{x}}\cdot[(N_z\Lambda)^{-1}+(\mathbf{A}_z)^{-1}]^{-1}\hat{\mathbf{x}}_z}
\end{aligned} \tag{6.18}$$

Having completed the μ integral, we are left with a complicated integral over Λ. To progress, we are forced to make some assumptions about the as yet arbitrary matrix $\mathbf{A}$. Our present more general assumptions lead to calculations that differ from the earlier work of e.g. [16, 82, 128]. An analytically tractable alternative is the transformation-invariant Jeffrey's prior which has the form $|\Sigma|^{-\frac{p+2}{2}}$ when the mean vector and covariance matrix are both unknown [146]. The predictive probability calculation is simpler than the current case but the sample covariance matrix is not regularized leading to ill-posedness when $N < p$. Our next question is for which choice(s) of $\mathbf{A}_z$ we can do also the integrals over Λ in (6.18) analytically. Expression (6.18), in line with [16, 82, 128], suggests using for the measure $p_z(\Lambda)$ over all positive definite matrices a Wishart distribution, which is of the form

$$p(\Lambda) = \frac{(\mathrm{Det}\Lambda)^{(r-p-1)/2}}{2^{rp/2}\Gamma_p(\frac{r}{2})(\mathrm{Det}\mathbf{S})^{r/2}}\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\mathbf{S}^{-1}\Lambda)} \tag{6.19}$$

Here the degrees of freedom, $r > p-1$, $\mathbf{S}$ is a positive definite symmetric $p \times p$ matrix (often termed the seed matrix), and the multivariate gamma function can be expressed in terms of the ordinary gamma function via:

$$\Gamma_p\left(\frac{r}{2}\right) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{r}{2} - \frac{j-1}{2}\right) \tag{6.20}$$

The Wishart distribution is unimodal and has an expectation $\mathbb{E}(\Lambda) = r\mathbf{S}$ (see Appendix A.2). The choice (6.19) is motivated solely by analytic tractability and implies that Λ is the empirical precision matrix of a set of r i.i.d. random vectors distributed as $\mathrm{N}_p(\mathbf{0}, \mathbf{S})$. Since (6.19) is normalised, for any positive definite $\mathbf{S}$, we can evaluate all integrals of the following form analytically:

$$\int_{\Xi^p} \mathrm{d}\Lambda \, (\mathrm{Det}\Lambda)^{(r-p-1)/2} \, \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}(\mathbf{S}^{-1}\Lambda)} = 2^{rp/2}\Gamma_p\left(\frac{r}{2}\right)(\mathrm{Det}\mathbf{S})^{r/2} \tag{6.21}$$

In order for (6.18) to acquire the form (6.21), we need a choice for $\mathbf{A}_z$ such that the following holds, for some $\gamma_0, \gamma_1 \in \mathbb{R}$: $\quad [(N_z\Lambda)^{-1} + (\mathbf{A}_z)^{-1}]^{-1} = \gamma_{1z}\Lambda + \gamma_{0z}\mathbb{I}$. Rewriting this condition gives:

$$\mathbf{A}_z(\gamma_{0z}, \gamma_1) = [(\gamma_{1z}\Lambda + \gamma_{0z}\mathbb{I})^{-1} - (N_z\Lambda)^{-1}]^{-1} \tag{6.22}$$

$\mathbf{A}_z$ has the same eigenvectors as Λ since $\mathbb{I}$ and Λ_z commute and a matrix has the same eigenvectors as its inverse. Denote the eigenvalues of Λ as $\{\lambda_\mu\}_{\mu=1}^{p}$. For $\mathbf{A}_z$ to be positive definite, we require all its eigenvalues to be positive which leads to restrictions on the domains of γ_{0z} and γ_{1z}. Each eigenvalue $\lambda \geq 0$ of Λ would give a corresponding eigenvalue $a(\lambda, z)$ for $\mathbf{A}_z$:

$$a(\lambda, z) = \frac{N_z\lambda(\gamma_{1z}\lambda + \gamma_{0z})}{(N_z - \gamma_{1z})\lambda - \gamma_{0z}} \tag{6.23}$$

We note that the zeros of $a(\lambda, z)$ occur at $\lambda \in \{-\gamma_{0z}/\gamma_{1z}, 0\}$, and that

$$\lambda \to 0: \;\; a(\lambda, z) = -N_z\lambda + \mathscr{O}(\lambda^2), \qquad \lambda \to \infty: \;\; a(\lambda, z) \approx \frac{N_z\gamma_{1z}\lambda}{N_z - \gamma_{1z}} \tag{6.24}$$

Labelling the smallest eigenvalue of Λ as $\lambda_{\min}$, the natural choice is to take

$$\gamma_{1z} \in (0, N_z], \qquad \gamma_{0z} \in (-\lambda_{\min}\gamma_{1z}, \lambda_{\min}(N_z - \gamma_{1z})) \tag{6.25}$$

This ensures that always $a(\lambda, z) > 0$ which is necessary since $\mathbf{A}$ should be invertible. We expect that $\lambda_{\min}$ will increase monotonically with $r - p$. This can be seen by considering the lower bound of the Marčenko-Pastur equation (1.15) i.e. $(1 - \sqrt{r/p})^2$. Upon making the

above choice for $\mathbf{A}_z$ and using (6.21), we would obtain for the integral (6.18):

$$\Psi_z = \mathrm{e}^{-\frac{1}{2}\gamma_{0z}\hat{\mathbf{x}}_z^2} \int \mathrm{d}\Lambda\; p_z(\Lambda) \frac{\mathrm{e}^{-\frac{1}{2}N_z \mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda) - \frac{1}{2}\gamma_{1z}\hat{\mathbf{x}}_z\cdot\Lambda\hat{\mathbf{x}}_z}}{\sqrt{\mathrm{Det}[N_z\Lambda^{1-N_z}(\gamma_{1z}\Lambda + \gamma_{0z}\mathbb{I})^{-1}]}} \tag{6.26}$$

where $\hat{\mathbf{x}}_z^2 = \hat{\mathbf{x}}_z \cdot \hat{\mathbf{x}}_z = \sum_{\nu=1}^{p} (\hat{x}_\nu^z)^2$. We can now evaluate (6.26) analytically, using (6.21), provided we choose for $p_z(\Lambda)$ the Wishart measure, and with either $\gamma_{0z} \to 0$ and $\gamma_{1z} \in (0, N_z)$, or with $\gamma_{1z} \to 0$ and $\gamma_{0z} \in (0, N_z\lambda_{\min})$. Alternative choices for $(\gamma_{0z}, \gamma_{1z})$ would lead to more complicated integrals than the Wishart one.

The two remaining analytically integrable candidate model branches imply the following choices for the inverse correlation matrix $\mathbf{A}_z$ of the prior $p_z(\mu|\mathbf{A}_z)$ for the class centres using (6.22):

$$\gamma_{0z} = 0: \quad \mathbf{A}_z = \frac{N_z\gamma_{1z}}{N_z - \gamma_{1z}}\Lambda, \qquad \gamma_{1z} = 0: \quad \mathbf{A}_z = \left[\gamma_{0z}^{-1}\,\mathbb{I} - (N_z\Lambda)^{-1}\right]^{-1} \tag{6.27}$$

Note that the case $\mathbf{A}_z \to 0$, a non-informative prior[1] for class means as in [16], corresponds to $(\gamma_{0z}, \gamma_{1z}) = (0,0)$. However, the two limits $\gamma_{0z} \to 0$ and $\gamma_{1z} \to 0$ will generally not commute, which can be inferred from working out (6.26) for the two special cases $\gamma_{0z} = 0$ and $\gamma_{1z} = 0$:

$$\gamma_{0z} = 0: \qquad \Psi_z = \left(\frac{\gamma_{1z}}{N_z}\right)^{\frac{p}{2}} \int \mathrm{d}\Lambda\; p_z(\Lambda)[\mathrm{Det}(\Lambda)]^{\frac{N_z}{2}} \mathrm{e}^{-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda) - \frac{1}{2}\gamma_{1z}\hat{\mathbf{x}}_z\cdot\Lambda\hat{\mathbf{x}}_z} \tag{6.28a}$$

$$\gamma_{1z} = 0: \qquad \Psi_z = \left(\frac{\gamma_{0z}}{N_z}\right)^{\frac{p}{2}} \mathrm{e}^{-\frac{1}{2}\gamma_{0z}\hat{\mathbf{x}}_z^2} \int \mathrm{d}\Lambda\; p_z(\Lambda)[\mathrm{Det}(\Lambda)]^{\frac{N_z-1}{2}} \mathrm{e}^{-\frac{1}{2}N_z\mathrm{Tr}(\hat{\mathbf{C}}_z\Lambda)} \tag{6.28b}$$

We can see that $\lim \gamma_{1z} \to 0$ of (6.28a) is not equivalent to $\lim \gamma_{0z} \to 0$ of (6.28b). This non-uniqueness of the limit $\mathbf{A}_z \to 0$ is found upon having done the integral over μ first.

6.3 The integrable model branches

We started with a rather general model family, and found that the requirement to do the integrals over both class centres and class covariance matrices analytically forces us for each class to set either γ_{0z} or γ_{1z} to zero (or both). These two remaining analytically integrable candidate model branches imply the forms of the inverse correlation matrix $\mathbf{A}_z$ of the prior $p_z(\mu|\mathbf{A}_z)$ for each class z given by (6.27). The following sections will examine these two cases in detail. The schematic in Figure 6.1 shows the relationship between the

[1] recall $\mathbf{A}_z$ is a precision matrix – the inverse of a covariance matrix

model branches characterized by the hyperparameter pair $(\gamma_{0z}, \gamma_{1z})$. We note the optimal hyperparameter choices $(\hat{\gamma}_{0z}, \hat{\gamma}_{1z})$ subsequently found in (6.35),(6.40).

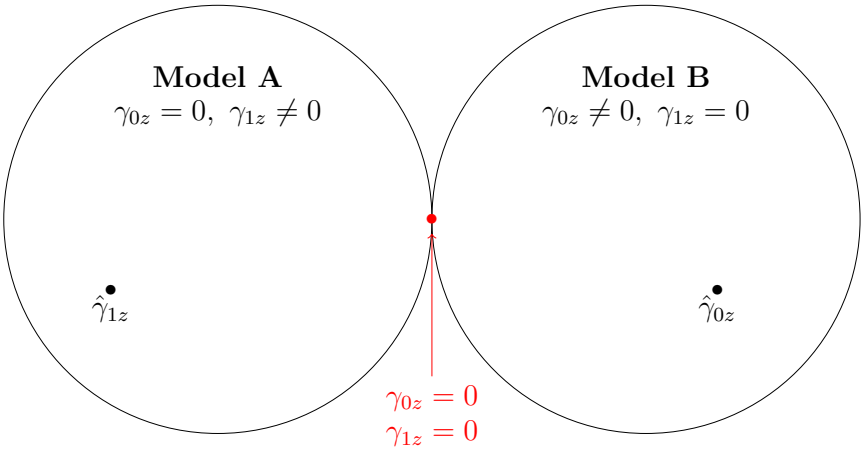

Fig. 6.1 Schematic diagram of two model branches. The common point (red) is at $(\gamma_{0z}, \gamma_{1z}) = (0,0)$ which corresponds to early literature [58, 82]. The optimized hyperparameters will produce more accurate classifiers than $(0,0)$ since $\Omega(\mathscr{H}, N, \mathscr{D})$ is minimized.

6.3.1 The case $\gamma_{0z} = 0$: model A

We now choose $\gamma_{0z} = 0$, and substitute for each $z = \{1 \ldots C\}$, the Wishart distribution (6.19) into (6.26), with seed matrix $\mathbf{S} = k_z \mathbb{I}$. This choice is named Quadratic Bayes in [16]. We also define the $p \times p$ matrix $\hat{\mathbf{M}}_z$ with entries

$$\hat{M}^z_{\mu\nu} = \hat{\mathbf{x}}^z_\mu \hat{\mathbf{x}}^z_\nu \tag{6.29}$$

The result of working out (6.26) is, using (6.21):

$$\Psi_z = \left(\frac{2^{N_z}\gamma_{1z}}{N_z k_z^{r_z}}\right)^{\frac{p}{2}} \frac{\Gamma_p(\frac{r_z+N_z}{2})}{\Gamma_p(\frac{r_z}{2})} [\mathrm{Det}(N_z\hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I})]^{-(r_z+N_z)/2} \tag{6.30}$$

This, in turn, allows us to evaluate (6.13):

$$\begin{aligned}\Omega(\mathscr{H}, N, \mathscr{D}) = &\frac{1}{2}Np\log(\pi) - \sum_{z=1}^{C} N_z \log p_z - \frac{1}{2}p\sum_{z=1}^{C}\left[\log(\gamma_{1z}/N_z) - r_z \log k_z\right] \\ &- \sum_{z=1}^{C} \log\left[\frac{\Gamma_p(\frac{r_z+N_z}{2})}{\Gamma_p(\frac{r_z}{2})}\right] + \frac{1}{2}\sum_{z=1}^{C}(r_z+N_z)\log\mathrm{Det}(N_z\hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I})\end{aligned} \tag{6.31}$$

The hyperparameters of our problem are $\{p_z, \gamma_{1z}, r_z, k_z\}$, for $z = \{1,\ldots,C\}$. If we choose flat hyper-priors, to close the Bayesian inference hierarchy, their optimal values are obtained by minimizing (6.31), subject to the constraints $\sum_{z=1}^{C} p_z = 1$, $p_z \geq 0$ (rules of probability), $r_z \geq p$ (Wishart distribution definition Appendix A.2), $\gamma_{1z} \in [0, N_z]$ (for positive definite $\mathbf{A}$), and $k_z > 0$ (for positive definite $\mathbf{S} = k_z \mathbb{I}$). We now work out the relevant extremization equations, using the general identity $\partial_x \log \mathrm{Det}\mathbf{Q} = \mathrm{Tr}(\mathbf{Q}^{-1}\partial_x \mathbf{Q})$:

- Minimization over p_z: Using the constraint of probability normalisation

$$\begin{aligned} 0 &= \frac{\partial}{\partial p_z}\left\{ \sum_{z'=1}^{C} N_{z'} \log p_{z'} - \lambda \Big(\sum_{z'=1}^{c} p_{z'} - 1 \Big) \right\} = \frac{N_z}{p_z} - \lambda \\ 1 &= \sum_{z=1}^{c} \frac{N_z}{\lambda} \quad \Rightarrow \quad \lambda = N \end{aligned} \tag{6.32}$$

  As expected for the generative model, the hyperparameter estimate is $\hat{p}_z = N_z/N$ i.e. the empirical fraction of samples in class z.

- Minimization over k_z:

$$k_z = 0 \text{ or } r_z = N_z \left[\frac{pk_z}{\mathrm{Tr}(N_z \hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I})^{-1}} - 1 \right]^{-1} \tag{6.33}$$

- Minimization over r_z, using the expression for the multivariate gamma function (6.20) and the digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$:

$$\begin{aligned} r_z = p \text{ or } \log k_z = \frac{1}{p}\sum_{j=1}^{p} \left[\psi\left(\frac{r_z + N_z - j + 1}{2} \right) - \psi\left(\frac{r_z - j + 1}{2} \right) \right] \\ - \frac{1}{p} \log \mathrm{Det}(N_z \hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I}) \end{aligned} \tag{6.34}$$

- Minimization over γ_{1z}:

$$\gamma_{1z} \in \{0, N_z\} \text{ or } \gamma_{1z} = \frac{1}{r_z + N_z}\left[\frac{1}{p} \mathrm{Tr}[(N_z \hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I})^{-1}\hat{\mathbf{M}}_z] \right]^{-1} \tag{6.35}$$

In addition we still need to satisfy the inequalities $r_z \geq p$, $\gamma_{1z} \in [0, N_z]$, and $k_z > 0$. The eigenvalues of the empirical covariance matrix $\hat{\mathbf{C}}_z$ are calculated once. However, we observe in the above results that, unless we choose $\gamma_{1z} = 0 \Rightarrow \mathbf{A} = 0$ or $\gamma_{1z} = N_z \Rightarrow \mathbf{A}^{-1} = 0$, we would during any iterative algorithmic solution of (6.33)-(6.35) have to invert the $p \times p$ matrix $N_z \hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + k_z^{-1}\mathbb{I}$ at each iteration step. This would be prohibitively slow, even with

the most efficient numerical diagonalization methods. We now consider the two plausible hyperparameter settings for γ_{1z} given $\gamma_{0z} = 0$:

$\gamma_{1z} = N_z$ *case:* From the assumed form of $\mathbf{A}$ (6.22), this hyperparameter setting leads to $\mathbf{A}^{-1} = \mathbf{0}$. The μ prior, defined in (6.17), becomes a delta function and forces all class centres to be in the origin.

$\gamma_{1z} = 0$ *case:* The precision matrix $\mathbf{A} = \mathbf{0}$. Therefore we are only left with the option $\gamma_{1z} \to 0$, corresponding to a flat prior for the class z.

We thereby arrive at the Quadratic Bayes model of [16], with hyperparameter formulae based on evidence maximization. It may be entirely possible that iterating with the full expression for γ_{1z} in (6.35) maybe produce superior classification results if computational time was no object.

6.3.2 The case $\gamma_{1z} = 0$: model B

We next inspect the alternative model branch by choosing $\gamma_{1z} = 0$, again substituting for each $z = \{1,\ldots,C\}$, the Wishart distribution (6.19) into (6.26) with seed matrix $\mathbf{S} = k_z \mathbb{I}$. The result is:

$$\Psi_z = \Big(\frac{2^{N_z-1}\gamma_{0z}}{N_z k_z^{r_z}}\Big)^{\frac{p}{2}} \frac{\Gamma_p(\frac{r_z+N_z-1}{2})}{\Gamma_p(\frac{r_z}{2})} [\mathrm{Det}(N_z\hat{\mathbf{C}}_z + k_z^{-1}\mathbb{I})]^{-(r_z+N_z-1)/2} \mathrm{e}^{-\frac{1}{2}\gamma_{0z}\hat{\mathbf{x}}_z^2} \tag{6.36}$$

For the quantity (6.13) we thereby find:

$$\begin{aligned}\Omega(\mathscr{H}, N, \mathscr{D}) = \frac{1}{2}Np\log(\pi) + \frac{1}{2}pC\log 2 - \sum_{z=1}^{C} N_z \log p_z - \frac{1}{2}p\sum_{z=1}^{C}\Big[\log\Big(\frac{\gamma_{0z}}{N_z}\Big) - r_z \log k_z\Big] \\ - \sum_{z=1}^{C}\log\Big[\frac{\Gamma_p(\frac{r_z+N_z-1}{2})}{\Gamma_p(\frac{r_z}{2})}\Big] + \frac{1}{2}\sum_{z=1}^{C}\gamma_{0z}\hat{\mathbf{x}}_z^2 + \frac{1}{2}\sum_{z=1}^{C}(r_z+N_z-1)\log\mathrm{Det}(N_z\hat{\mathbf{C}}_z + k_z^{-1}\mathbb{I})\end{aligned} \tag{6.37}$$

If as before we choose flat hyper-priors, the Bayes-optimal hyperparameters $\{p_z, \gamma_{1z}, r_z, k_z\}$, for $z = \{1,\ldots,C\}$ are found by minimizing $\Omega(\mathscr{H}, N, \mathscr{D})$ in (6.37), which is equivalent to maximizing the evidence in (6.7), subject to the constraints $\sum_{z=1}^{C} p_z = 1$, $p_z \geq 0$, $r_z \geq p$, $\gamma_{0z} \geq 0$, and $k_z > 0$. For the present model branch B, differentiation gives

- Minimization over p_z: $\hat{p}_z = N_z/N$.

- Minimization over k_z:

$$k_z = 0 \text{ or } r_z = (N_z - 1)\left[\frac{pk_z}{\mathrm{Tr}[(N_z\hat{\mathbf{C}}_z + k_z^{-1}\,\mathbb{I})^{-1}} - 1\right]^{-1} \tag{6.38}$$

- Minimization over r_z:

$$r_z = p \text{ or } \log k_z = \frac{1}{p}\sum_{j=1}^{p}\left[\psi\left(\frac{r_z + N_z - j}{2}\right) - \psi\left(\frac{r_z - j + 1}{2}\right)\right] - \frac{1}{p}\log\mathrm{Det}(N_z\hat{\mathbf{C}}_z + k_z^{-1}\,\mathbb{I}) \tag{6.39}$$

  Up to now, these equations are similar to (6.33),(6.34) apart from the factor $N_z - 1$. Minimization with respect to the prior-specific hyperparameter is different.

- Minimization over γ_{0z}:

$$\gamma_{0z} = p/\hat{\mathbf{x}}_z^2 \tag{6.40}$$

In addition we still need to satisfy the inequalities $r_z \geq p$ and $k_z > 0$. In contrast to the first integrable model branch A, here we are able to optimise over γ_{0z} without problems, and the resulting model B is distinct from the Quadratic Bayes classifier of [16] and appears to be novel.

6.3.3 Comparison of the two integrable model branches

Our initial family of models was parametrized by $(\gamma_{0z}, \gamma_{1z})$. We then found that the following two branches are analytically integrable, using Wishart priors for class-specific precision matrices:

$$\text{Model A}: \quad (\gamma_{0z}, \gamma_{1z}) = (0, \hat{\gamma}_{1z}) \;\text{ with }\; \hat{\gamma}_{1z} \to 0 \tag{6.41a}$$

$$\text{Model B}: \quad (\gamma_{0z}, \gamma_{1z}) = (\hat{\gamma}_{0z}, 0) \;\text{ with }\; \hat{\gamma}_{0z} \to p/\hat{\mathbf{x}}_z^2 \tag{6.41b}$$

Where conventional methods tend to determine hyperparameters via cross-validation, which is computationally expensive, here we optimize hyperparameters via evidence maximization. As expected, both models give $\hat{p}_z = N_z/N$. The hyperparameters (k_z, r_z) are to be solved from the following equations, in which $\rho_z(\xi)$ denotes the eigenvalue distribution of the empirical

covariance matrix $\hat{\mathbf{C}}_z$:

$$\text{A}: \quad k_z = 0 \quad \text{or} \quad r_z = N_z\Big[\frac{1}{\int \mathrm{d}\xi\ \rho_z(\xi)(N_z k_z \xi + 1)^{-1}} - 1\Big]^{-1} \tag{6.42a}$$

$$r_z = p \quad \text{or} \quad \frac{1}{p}\sum_{j=1}^{p}\Big[\psi\Big(\frac{r_z+N_z-j+1}{2}\Big) - \psi\Big(\frac{r_z-j+1}{2}\Big)\Big] = \int \mathrm{d}\xi\ \rho_z(\xi)\log(N_z k_z \xi + 1) \tag{6.42b}$$

$$\text{B}: \quad k_z = 0 \quad \text{or} \quad r_z = (N_z-1)\Big[\frac{1}{\int \mathrm{d}\xi\ \rho_z(\xi)(N_z k_z \xi + 1)^{-1}} - 1\Big]^{-1} \tag{6.42c}$$

$$r_z = p \quad \text{or} \quad \frac{1}{p}\sum_{j=1}^{p}\Big[\psi\Big(\frac{r_z+N_z-j}{2}\Big) - \psi\Big(\frac{r_z-j+1}{2}\Big)\Big] = \int \mathrm{d}\xi\ \rho_z(\xi)\log(N_z k_z \xi + 1) \tag{6.42d}$$

We see that the equations for (k_z, r_z) of models A and B differ only in having the replacement $N_z \to N_z - 1$ in certain places. Hence we will have $(k_z^A, r_z^A) = (k_z^B, r_z^B) + \mathscr{O}(N_z^{-1})$. Hyperparameter r_z is a monotonically decreasing function of k_z. This is highlighted, along with the boundary condition imposed by the Wishart distribution $r_z > p$, in Figure 6.2.

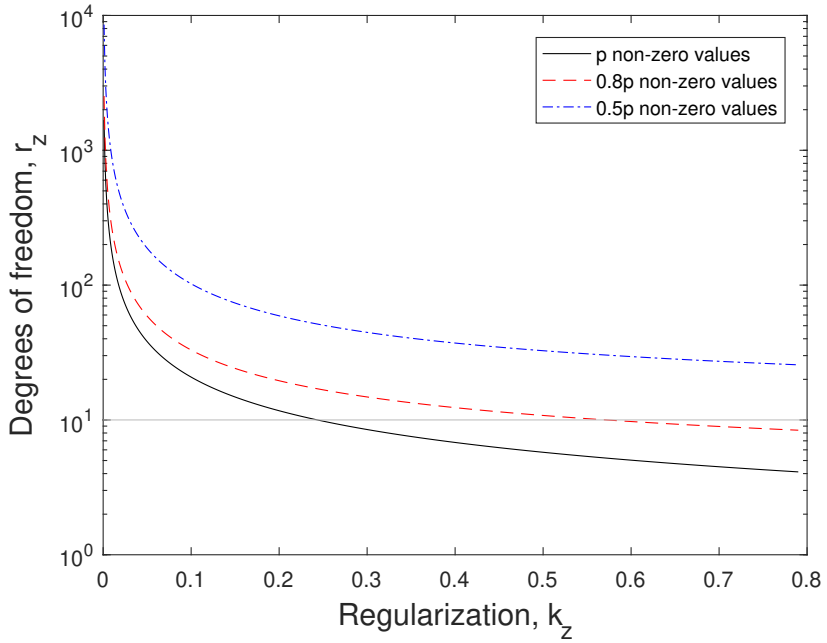

Fig. 6.2 Plot of hyperparameters r_z and k_z in the evidence maximization with $N = 10, p = 10$ and varying proportions of positive (non-zero) eigenvalues. The k_z value where the curves cross the dotted $r_z = p$ line represents the maximum value of k_z i.e. $k_{\max,z}$. Values relate to model A.

The most probable model in a Bayesian sense. Recalling that $\Omega(\mathscr{H}, N, \mathscr{D})$ represents the negative log posterior (6.9), the most plausible model, given the data, is the one that gives the smallest value for $\Omega(\mathscr{H}, N, \mathscr{D})$. We compute the difference between this quantity for the two branches by noting that

$$\Omega(\mathscr{H}, N, \mathscr{D}) = \sum_{z=1}^{C} \Omega_z(\mathscr{H}, N, \mathscr{D}) \tag{6.43}$$

and using (6.31) and (6.37)

$$\begin{aligned}
\Delta_z &= \Omega_z(\mathscr{H}_A, N, \mathscr{D}) - \Omega_z(\mathscr{H}_B, N, \mathscr{D}) \\
&= -\frac{1}{2}p(1+\log 2) - \frac{1}{2}p\log(\hat{X}_z^2/p) \\
&+ \lim_{\hat{\gamma}_{1z}\to 0}\Big\{ -\frac{1}{2}p\left[\log\hat{\gamma}_{1z} + r_z^B\log k_z^B - r_z^A\log k_z^A\right] - \log\left[\frac{\Gamma_p(\frac{r_z^A+N_z}{2})}{\Gamma_p(\frac{r_z^A}{2})}\right] + \log\left[\frac{\Gamma_p(\frac{r_z^B+N_z-1}{2})}{\Gamma_p(\frac{r_z^B}{2})}\right] \\
&+ \frac{1}{2}(r_z^A+N_z)\log\mathrm{Det}(N_z\hat{\mathbf{C}}_z + \gamma_{1z}\hat{\mathbf{M}}_z + (1/k_z^A)\mathbb{I}) - \frac{1}{2}(r_z^B+N_z-1)\log\mathrm{Det}(N_z\hat{\mathbf{C}}_z + (1/k_z^B)\mathbb{I})\Big\} \\
&= \infty
\end{aligned} \tag{6.44}$$

where the hyperparameters remain finite for a given p and N. It follows that model B is always more plausible than model A, for any class z as long as there is no model mismatch i.e. the true-data generating model is Gaussian. We also note that, had we replaced the optimal value $\gamma_{0z} = p/\hat{X}_z^2$ in model B by the ad hoc choice $\gamma_{0z} = 0$ (which would have corresponded to $\mathbf{A} = 0$), we would again have arrived at a much less plausible model, for which the value of $\Omega_z(\mathscr{H}, N, \mathscr{D})$ relative to that of the optimal γ_{0z} would be increased by a diverging term $\frac{1}{2}p\log(1/\gamma_{0z})$.

6.3.4 Expressions for the predictive probability

We are finally in a position to derive the predictive probabilities for both models. Expressions for $\Omega(\mathscr{H}, N, \mathscr{D})$ were found (6.31)(6.37) to estimate the hyperparameters. However, for the predictive probability distribution, $p(y_0|\mathbf{x}_0, \mathscr{D})$ (6.10), we now require $\Omega(\mathscr{H}, N+1, \mathscr{D})$ which incorporates the additional test data sample. The following algebra carefully makes the replacements $N \to N+1$ and $N_z \to N_z + \delta_{y_0,z}$ (taking care not to change anything in the hyperparameter equations). First recall the definitions $I_z = \{i|\ y_i = z\}$ and $N_z = |I_z| = \sum_{i=1}^{N}\delta_{z,y_i}$.

The general form of the empirical mean and covariance matrix using $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is

$$\hat{x}^z_\mu = \frac{1}{N_z} \sum_{i \in I_z} x_{i\mu}, \qquad \hat{C}^z_{\mu\nu} = \frac{1}{N_z} \sum_{i \in I_z} (x_{i\mu} - \hat{x}^z_\mu)(x_{i\nu} - \hat{x}^z_\nu) \tag{6.45}$$

Modifying the empirical mean to correctly incorporate the new data point $(\mathbf{x}_0, y_0)$

$$\hat{x}^z_\mu = \begin{cases} \frac{1}{N_{y_0}+1} \left[\sum_{i=1}^N \delta_{y_0 y_i} x_{i\mu} + x_{0\mu} \right] & \text{if } z = y_0 \\ \hat{x}^z_\mu & \text{if } z \neq y_0 \end{cases} \tag{6.46}$$

and similarly for $\hat{C}^z_{\mu\nu}$. Re-writing as transformations in terms of delta functions:

$$\begin{aligned} \hat{x}^z_\mu &\to (1 - \delta_{z y_0}) \hat{x}^z_\mu + \frac{\delta_{z y_0}}{N_{y_0}+1} \left[\sum_{i=1}^N \delta_{y_0 y_i} x_{i\mu} + x_{0\mu} \right] \\ &= \hat{x}^z_\mu + \frac{\delta_{z y_0}}{N_{y_0}+1} \left[N_{y_0} \hat{x}^z_\mu + x_{0\mu} - (N_{y_0}+1) \hat{x}^z_\mu \right] \\ &= \hat{x}^z_\mu + \frac{\delta_{z y_0}}{N_{y_0}+1} \left(x_{0\mu} - \hat{x}^z_\mu \right) \end{aligned} \tag{6.47}$$

We can now replace the average term in the empirical covariance matrix $\hat{C}^z$ with (6.47) and treating $z = y_0$ and $z \neq y_0$ separately.

$$
\begin{aligned}
\hat{C}^z_{\mu\nu} \to\ & (1-\delta_{zy_0})\hat{C}^z_{\mu\nu} \\
& + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{ \sum_{i=1}^{N} \delta_{y_0 y_i}\Big[x_{i\mu} - \hat{x}^z_\mu - \frac{1}{N_{y_0}+1}\Big(x_{0\mu}-\hat{x}^z_\mu\Big)\Big]\Big[x_{i\nu} - \hat{x}^z_\nu - \frac{1}{N_{y_0}+1}\Big(x_{0\nu}-\hat{x}^z_\nu\Big)\Big] \\
& + \Big[x_{0\mu} - \hat{x}^z_\mu - \frac{1}{N_{y_0}+1}\Big(x_{0\mu}-\hat{x}^z_\mu\Big)\Big]\Big[x_{0\nu} - \hat{x}^z_\nu - \frac{1}{N_{y_0}+1}\Big(x_{0\nu}-\hat{x}^z_\nu\Big)\Big]\Big\} \\
\\
& = (1-\delta_{zy_0})\hat{C}^z_{\mu\nu} \\
& + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{ \sum_{i=1}^{N} \delta_{y_0 y_i}\Big[x_{i\mu} - \hat{x}^z_\mu\Big]\Big[x_{i\nu} - \hat{x}^z_\nu\Big] - \frac{2}{N_{y_0}+1}\sum_{i=1}^{N} \delta_{y_0 y_i}\Big[x_{i\mu} - \hat{x}^z_\mu\Big]\Big[x_{0\nu}-\hat{x}^z_\nu\Big] \\
& - \frac{2}{N_{y_0}+1}\sum_{i=1}^{N} \delta_{y_0 y_i}\Big[x_{i\nu} - \hat{x}^z_\nu\Big]\Big[x_{0\mu}-\hat{x}^z_\mu\Big] + \frac{1}{(N_{y_0}+1)^2}\sum_{i=1}^{N} \delta_{y_0 y_i}\Big[x_{0\mu}-\hat{x}^z_\mu\Big]\Big[x_{0\nu}-\hat{x}^z_\nu\Big] \\
& + \frac{N^2_{y_0}}{(N_{y_0}+1)^2}\Big[x_{0\mu}-\hat{x}^z_\mu\Big]\Big[x_{0\nu}-\hat{x}^z_\nu\Big]\Big\} \\
\\
& = (1-\delta_{zy_0})\hat{C}^z_{\mu\nu} \\
& + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{N_{y_0}\hat{C}^{y_0}_{\mu\nu} + \frac{N_{y_0}}{(N_{y_0}+1)^2}(x_{0\mu}-\hat{x}^z_\mu)(x_{0\nu}-\hat{x}^z_\nu) + \frac{N^2_{y_0}}{(N_{y_0}+1)^2}(x_{0\mu}-\hat{x}^z_\mu)(x_{0\nu}-\hat{x}^z_\nu)\Big\}
\end{aligned}
\tag{6.48}
$$

Inserting the definition for $\hat{C}^z_{\mu\nu}$ from (6.45)

$$
\begin{aligned}
\hat{C}^z_{\mu\nu} &= (1-\delta_{zy_0})\hat{C}^z_{\mu\nu} + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{N_{y_0}\hat{C}^{y_0}_{\mu\nu} + \frac{N_{y_0}}{N_{y_0}+1}(x_{0\mu}-\hat{x}^z_\mu)(x_{0\nu}-\hat{x}^z_\nu)\Big\} \\
&= \hat{C}^z_{\mu\nu} + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{N_{y_0}\hat{C}^{y_0}_{\mu\nu} - (N_{y_0}+1)\hat{C}^{y_0}_{\mu\nu} + \frac{N_{y_0}}{N_{y_0}+1}(x_{0\mu}-\hat{x}^z_\mu)(x_{0\nu}-\hat{x}^z_\nu)\Big\} \\
&= \hat{C}^z_{\mu\nu} + \frac{\delta_{zy_0}}{N_{y_0}+1}\Big\{\frac{N_{y_0}}{N_{y_0}+1}(x_{0\mu}-\hat{x}^z_\mu)(x_{0\nu}-\hat{x}^z_\nu) - \hat{C}^{y_0}_{\mu\nu}\Big\}
\end{aligned}
\tag{6.49}
$$

We introduce the $p \times p$ matrix $\mathbf{M}_{y_0}$, with entries

$$
M^{y_0}_{\mu\nu} = (x_{0\mu}-\hat{x}^{y_0}_\mu)(x_{0\nu}-\hat{x}^{y_0}_\nu) \tag{6.50}
$$

We find it more convenient to work with the difference $\Omega(\mathscr{H},N+1,\mathscr{D})-\Omega(\mathscr{H},N,\mathscr{D})$. Using (6.37) and (6.43), this leads us to

$$\begin{aligned}
&\Omega(\mathscr{H},N+1,\mathscr{D})-\Omega(\mathscr{H},N,\mathscr{D}) \;=\; \Omega_{y_0}(\mathscr{H},N+1,\mathscr{D})-\Omega_{y_0}(\mathscr{H},N,\mathscr{D})\\
&=\frac{1}{2}p\log(\pi)-\log p_{y_0}-\frac{1}{2}p\log\left(\frac{N_{y_0}}{N_{y_0}+1}\right)-\log\left[\frac{\Gamma_p\left(\frac{r_{y_0}+N_{y_0}}{2}\right)}{\Gamma_p\left(\frac{r_{y_0}+N_{y_0}-1}{2}\right)}\right]\\
&+\frac{1}{2}\gamma_{0y_0}\left[\frac{2}{N_{y_0}+1}\hat{\mathbf{x}}_{y_0}\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})+\frac{1}{(N_{y_0}+1)^2}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})^2\right]\\
&+\frac{1}{2}(r_{y_0}+N_{y_0})\log\mathrm{Det}\left(N_{y_0}\hat{\mathbf{C}}_{y_0}+k_{y_0}^{-1}\mathbb{I}+\frac{N_{y_0}}{N_{y_0}+1}\mathbf{M}_{y_0}\right)\\
&-\frac{1}{2}(r_{y_0}+N_{y_0}-1)\log\mathrm{Det}\left(N_{y_0}\hat{\mathbf{C}}_{y_0}+k_{y_0}^{-1}\mathbb{I}\right)\\
&=\frac{1}{2}p\log(\pi)-\log p_{y_0}-\frac{1}{2}p\log\left(\frac{N_{y_0}}{N_{y_0}+1}\right)-\log\left[\frac{\Gamma_p\left(\frac{r_{y_0}+N_{y_0}}{2}\right)}{\Gamma_p\left(\frac{r_{y_0}+N_{y_0}-1}{2}\right)}\right]\\
&+\frac{1}{2}\gamma_{0y_0}\left[\frac{2}{N_{y_0}+1}\hat{\mathbf{x}}_{y_0}\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})+\frac{1}{(N_{y_0}+1)^2}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})^2\right]\\
&+\frac{1}{2}(r_{y_0}+N_{y_0})\log\left[\frac{\mathrm{Det}(N_{y_0}\hat{\mathbf{C}}_{y_0}+k_{y_0}^{-1}\mathbb{I}+\frac{N_{y_0}}{N_{y_0}+1}\mathbf{M}_{y_0})}{\mathrm{Det}(N_{y_0}\hat{\mathbf{C}}_{y_0}+k_{y_0}^{-1}\mathbb{I})}\right]+\frac{1}{2}\log\mathrm{Det}(N_{y_0}\hat{\mathbf{C}}_{y_0}+k_{y_0}^{-1}\mathbb{I})
\end{aligned}\tag{6.51}$$

Introducing the short-hand $\Xi_z=N_z\hat{\mathbf{C}}_z+k_z^{-1}\mathbb{I}$

$$\begin{aligned}
&\Omega(\mathscr{H},N+1,\mathscr{D})-\Omega(\mathscr{H},N,\mathscr{D})\\
&=\frac{1}{2}p\log(\pi)-\log p_{y_0}-\frac{1}{2}p\log\left(\frac{N_{y_0}}{N_{y_0}+1}\right)-\log\left[\frac{\Gamma_p(\frac{r_{y_0}+N_{y_0}}{2})}{\Gamma_p(\frac{r_{y_0}+N_{y_0}-1}{2})}\right]\\
&+\frac{1}{2}\gamma_{0y_0}\left[\frac{2}{N_{y_0}+1}\hat{\mathbf{x}}_{y_0}\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})+\frac{1}{(N_{y_0}+1)^2}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})^2\right]\\
&+\frac{1}{2}\log\mathrm{Det}[\Xi_{y_0}]+\frac{1}{2}(r_{y_0}+N_{y_0})\log\mathrm{Det}\left[\mathbb{I}+\Xi_{y_0}^{-\frac{1}{2}}\frac{N_{y_0}\mathbf{M}_{y_0}}{N_{y_0}+1}\Xi_{y_0}^{-\frac{1}{2}}\right]
\end{aligned}\tag{6.52}$$

Note that Ξ_z and $\hat{\mathbf{C}}_z$ share the same eigenvector basis. Let us work out some of the terms in this expression further:

- The multivariate gamma functions:

$$\frac{\Gamma_p\left(\frac{r_{y_0}+N_{y_0}}{2}\right)}{\Gamma_p(\frac{r_{y_0}+N_{y_0}-1}{2})}=\prod_{j=1}^{p}\frac{\Gamma\left(\frac{r_{y_0}+N_{y_0}-j+1}{2}\right)}{\Gamma\left(\frac{r_{y_0}+N_{y_0}-j}{2}\right)}=\frac{\Gamma\left(\frac{r_{y_0}+N_{y_0}}{2}\right)}{\Gamma\left(\frac{r_{y_0}+N_{y_0}-p}{2}\right)} \tag{6.53}$$

- We calculate the determinant involving $\mathbf{M}_{y_0}$ by considering the definition $\mathbf{M}_{y_0} = (\mathbf{x}_0-\hat{\mathbf{x}}^{y_0})(\mathbf{x}_0-\hat{\mathbf{x}}^{y_0})^T$ from (6.50). The matrix is proportional[2] to a projection matrix and therefore has only one nonzero eigenvalue λ_{y_0}. We can compute this nontrivial eigenvalue simply via

$$\begin{aligned}\lambda_{y_0} &= \mathrm{Tr}[\Xi_{y_0}^{-\frac{1}{2}}\mathbf{M}_{y_0}\Xi_{y_0}^{-\frac{1}{2}}] \;=\; \mathrm{Tr}[\Xi_{y_0}^{-1}\mathbf{M}_{y_0}] \\ &= \sum_{\mu,\nu=1}^{p}(\Xi_{y_0}^{-1})_{\mu\nu}(x_{0\mu}-\hat{x}_\mu^{y_0})(x_{0\nu}-\hat{x}_\nu^{y_0}) = (\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\cdot\Xi_{y_0}^{-1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\end{aligned} \tag{6.54}$$

The Mahalanobis distance $(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\cdot\Xi_{y_0}^{-1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})$ represents the distance between the new data point, $\mathbf{x}_0$, and the mean class $\hat{\mathbf{x}}_{y_0}$. It is a positive value since Ξ_{y_0} is positive definite. Hence

$$\begin{aligned}&\Omega(\mathscr{H},N+1,\mathscr{D}) = \\ &\Omega(\mathscr{H},N,\mathscr{D})+\frac{1}{2}p\log(\pi)-\log p_{y_0}-\frac{1}{2}p\log\left(\frac{N_{y_0}}{N_{y_0}+1}\right)-\log\left[\frac{\Gamma\left(\frac{r_{y_0}+N_{y_0}}{2}\right)}{\Gamma\left(\frac{r_{y_0}+N_{y_0}-p}{2}\right)}\right] \\ &+\frac{1}{2}\gamma_{0y_0}\left[\frac{2}{N_{y_0}+1}\hat{\mathbf{x}}_{y_0}\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})+\frac{1}{(N_{y_0}+1)^2}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})^2\right] \\ &+\frac{1}{2}\log\mathrm{Det}[\Xi_{y_0}]+\frac{1}{2}(r_{y_0}+N_{y_0})\log\left[1+\frac{N_{y_0}}{N_{y_0}+1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\cdot\Xi_{y_0}^{-1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\right]\end{aligned} \tag{6.55}$$

This then leads to

$$p(y_0|x_0,\mathscr{D}) = \frac{\mathrm{e}^{-\Omega(\mathscr{H},N+1,\mathscr{D})}}{\sum_{y_0'=1}^{C}\mathrm{e}^{-\Omega(\mathscr{H},N+1,\mathscr{D})}\big|_{y_0=y_0'}} \tag{6.56}$$

[2]For matrix $\mathbf{M}=\mathbf{x}\mathbf{x}^T$ defined as the outer product of unnormalized vector $\mathbf{x}$, $\mathbf{M}^2=(\mathbf{x}\mathbf{x}^T)(\mathbf{x}\mathbf{x}^T)=\mathbf{x}(\mathbf{x}^T\mathbf{x})\mathbf{x}^T\propto\mathbf{M}$ since $\mathbf{x}^T\mathbf{x}$ is a scalar value. By construction, all columns of $\mathbf{M}$ are proportional to each other so the matrix is rank one.

The predictive probability for model B:

$$p(y_0|\mathbf{x}_0,\mathscr{D}) = \frac{W_{y_0}\mathrm{e}^{-\frac{\gamma_{0y_0}}{2(N_{y_0}+1)}\left[2\hat{\mathbf{x}}_{y_0}\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})+\frac{1}{N_{y_0}+1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})^2\right]}\left(1+\frac{N_{y_0}}{N_{y_0}+1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\cdot\Xi_{y_0}^{-1}(\mathbf{x}_0-\hat{\mathbf{x}}_{y_0})\right)^{-\frac{1}{2}(r_{y_0}+N_{y_0})}}{\sum_{z=1}^{C}W_z\mathrm{e}^{-\frac{\gamma_{0z}}{2(N_z+1)}\left[2\hat{\mathbf{x}}_z\cdot(\mathbf{x}_0-\hat{\mathbf{x}}_z)+\frac{1}{N_z+1}(\mathbf{x}_0-\hat{\mathbf{x}}_z)^2\right]}\left(1+\frac{N_z}{N_z+1}(\mathbf{x}_0-\hat{\mathbf{x}}_z)\cdot\Xi_z^{-1}(\mathbf{x}_0-\hat{\mathbf{x}}_z)\right)^{-\frac{1}{2}(r_z+N_z)}} \tag{6.57}$$

with $p_z = N_z/N$, and

$$W_z = p_z\left(\frac{N_z}{N_z+1}\right)^{\frac{p}{2}}\frac{\Gamma\left(\frac{r_z+N_z}{2}\right)}{\Gamma\left(\frac{r_z+N_z-p}{2}\right)}[\mathrm{Det}\Xi_z]^{-\frac{1}{2}}, \tag{6.58}$$

$$\gamma_{0z} = p/\hat{\mathbf{x}}_z^2 \qquad \text{and} \qquad \Xi_z = \mathrm{N_z}\hat{\mathbf{C}}_z+\mathrm{k_z^{-1}}\mathbb{I}$$

Upon repeating the same calculations for model A one finds that its predictive probability is obtained from expression (6.57) simply by setting γ_{0y_0} to zero (keeping in mind that for model A we would also insert into this formula distinct values for the optimal hyperparameters k_z and r_z). In this case, our predictive probability simplifies to the multivariate student-t form of earlier work such as [82]. The main difference being our method of estimating hyperparameters through evidence maximization via (6.42a)-(6.42d). Before interpreting (6.57), it is worthwhile recapping the assumptions underlying its derivation with links to the relevant equations.

- Generative Bayesian classifier with multivariate Gaussian class-conditional probability distributions (6.8).
- A Wishart prior for the precision matrix $p(\Lambda|r,\mathbf{S})$ (6.19)
- Quadratic Bayes [16] assumption for the seed matrix of the Wishart prior i.e. $\mathbf{S} = k\mathbb{I}$.
- A multivariate Gaussian prior for class means $p_z(\mu|\mathbf{A}_z)$ (6.17).
- Assumptions on the structure of the hyperparameter $\mathbf{A}_z$ (6.22).

We hope made all assumptions and calculations behind the hyperparameter estimation and the predictive probability expression have been made explicit. This should enable us to re-examine one or more of them in future work in order to improve classification accuracy.

Interpretation of the predictive probability (6.57). The regularized covariance matrix Ξ_z is non-singular when $N_z < p$ since $k_z > 0$ (unlike the unregularized case) allowing classification of high-dimensional datasets. Large values of hyperparameter k_z (low regularization), occur when there is sufficient data. Conversely small k_z (large regularization) occur when there is insufficient data.

As expected, a smaller Mahalanobis distance produces a larger contribution from that class. Examining the term containing this distance in the numerator, it can be seen that increasing k_z has the effect of reducing $p(y_0|\mathbf{x}_0, \mathscr{D}, \mathscr{H})$.

We now proceed to examine our key equation (6.57) and the associated hyperparameter estimation by examining results on synthetic, sanitized and real experimental data.

6.4 Phenomenology of the classifiers

6.4.1 Hyperparameters: LOOCV versus evidence maximization

In statistical inference, hyperparameters need to be determined to find the predictive probability. Yet the literature is full of comparisons between existing and new classification methods with little attention paid to describing this important step. Methods may be different between classifiers or not fully described leading to unfair comparisons. Therefore accurately specifying the protocol used is essential for reproducibility.

The most commonly used measure of classification performance is the percentage of samples correctly predicted on unseen data (equivalently, the trace of the confusion matrix, see Appendix A.7), and most Bayesian classification methods also use this measure as the optimization target for hyperparameters, via cross-validation. To make this precise, let the classification function, $f(\mathbf{x}|\mathscr{D}, \mathscr{H}) : \mathbf{x} \to y$ where $\mathbf{x} \in \mathbb{R}^p$, $y \in \{1, \ldots, C\}$. Then the classification accuracy is

$$\mathbb{E}[\ \mathrm{I}(\mathrm{f}(\mathbf{x}|\mathscr{D}, \mathscr{H}) = \mathrm{y})\ ] \tag{6.59}$$

where the expectation is over the joint distribution $p(\mathbf{x}, y)$ and $\mathrm{I}(.)$ represents the indicator function. Instead, our method of hyperparameter optimization maximizes the evidence term in the Bayesian inference (6.7). In k-fold cross-validation one needs to diagonalize for each outcome class a $p \times p$ matrix k times, whereas using the evidence maximization route one needs to diagonalize such matrices only once, giving a factor k reduction in what for large p is

the dominant contribution to the numerical demands. Moreover, cross-validation introduces fluctuations into the hyperparameter computation (via the random separations into training and validation sets), whereas evidence maximization is strictly deterministic.

The two routes, cross-validation versus evidence maximization, need not necessarily lead to coincident hyperparameter estimates. In order to investigate such possible differences we generated synthetic datasets with equal class sizes $N_1 = N_2 = 50$, and with input vectors of dimension $p = 50$. For numerical purposes, it was important to specific the maximum value, $k_{\max,z}$ as either the upper limit defined by the condition $r_z > p - 1$ (if such a limit exists, dependent on the data realization), or otherwise set numerically to a fixed large value (see Figure 6.2). To be clear, we describe the two different methods of arriving at the classification accuracy.

Leave-one-out cross-validation (LOOCV). Using a 100×100 grid of values for the hyperparameters k_1 and k_2, with $k_z \in [0, k_{\max,z}]$, we calculated the LOOCV estimator of classification accuracy for unseen cases, for a single data realisation. The values of (r_1, r_2) were determined via evidence maximization, using formula (6.42b) (i.e. following model branch A, with the noninformative prior for the class means). The location of the maximum of the resulting surface determines the LOOCV estimate of the optimal hyperparameters (k_1, k_2).

Evidence maximization. The optimized hyperparameters (k_1, k_2) are found via the evidence maximization method of (6.42a)-(6.42d).

Figure 6.3 shows the resulting surface for uncorrelated data, i.e. $\Sigma_1 = \Sigma_2 = \mathbb{I}$ (linear discriminant analysis). The comparison points from our evidence-based optimal hyperparameters (k_1, k_2) are shown in Table 6.1. The small values for (k_1, k_2) imply that the model correctly infers that the components of $\mathbf{x}$ in each class are most likely uncorrelated. The same protocol was subsequently repeated for correlated data, using a Toeplitz covariance matrix defined by its first row $(p, p-1, \ldots, 2, 1)$, the results of which are shown in Figure 6.4 and Table 6.1. The larger values for (k_1, k_2) imply that here the model correctly infers that the components of $\mathbf{x}$ in each class are correlated. In both cases the differences between optimal hyperparameter values defined via LOOCV as opposed to evidence maximization are seen to be minor.

(k_1, k_2)	*method*	*model A*	*model B*
Uncorrelated data	Cross-validation	(1-5%, 1-10%)	(1%, 1%)
	Evidence maximization	(3%, 2%)	(2%, 1%)
Correlated data	Cross-validation	(86-92%, 55-60%)	(56-99%, 47-94%)
	Evidence maximization	(93%, 94%)	(92%, 93%)

Table 6.1 Comparison of hyperparameter estimation using cross-validation and evidence maximization for correlated and uncorrelated data. Entries are the values of (k_1, k_2), given as a percentage of each class k_{max}, corresponding to the maximum classification accuracy (within the granularity of our numerical experiments). A range of values is given when they all share the same classification accuracy.

Classification accuracy (%)	*method*	*model A*	*model B*
Uncorrelated data	Cross-validation	87%	86%
	Evidence maximization	87%	85%
Correlated data	Cross-validation	69%	63%
	Evidence maximization	64%	62%

Table 6.2 Comparison of classification accuracy using cross-validation and evidence maximization methods for estimating hyperparameters using the same data as Figures 6.3, 6.4.

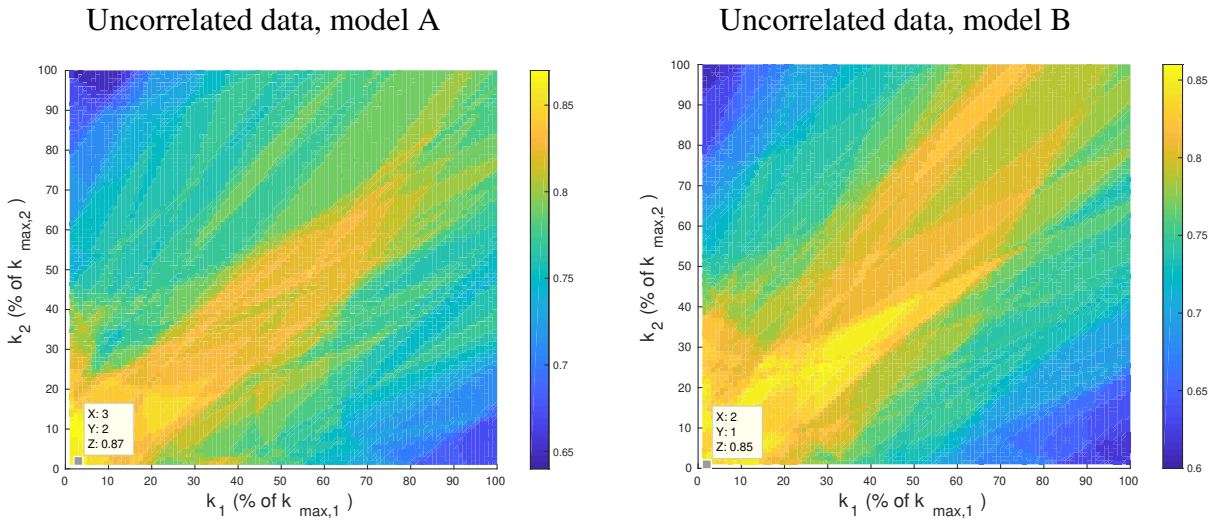

Fig. 6.3 LOOCV classification accuracy in (k_1, k_2) space for uncorrelated synthetic data, with class means $\mu_1 = (0,0,\ldots,0)$ and $\mu_2 = (2.5,0,\ldots,0)$, population covariance matrices $\Sigma_1 = \Sigma_2 = \mathbb{I}$, and covariate dimension $p = 50$. The hyperparameters for models A (left) and B (right) were determined via (6.42a)-(6.42d) and labelled on the plot. The results are based on a single data realization.

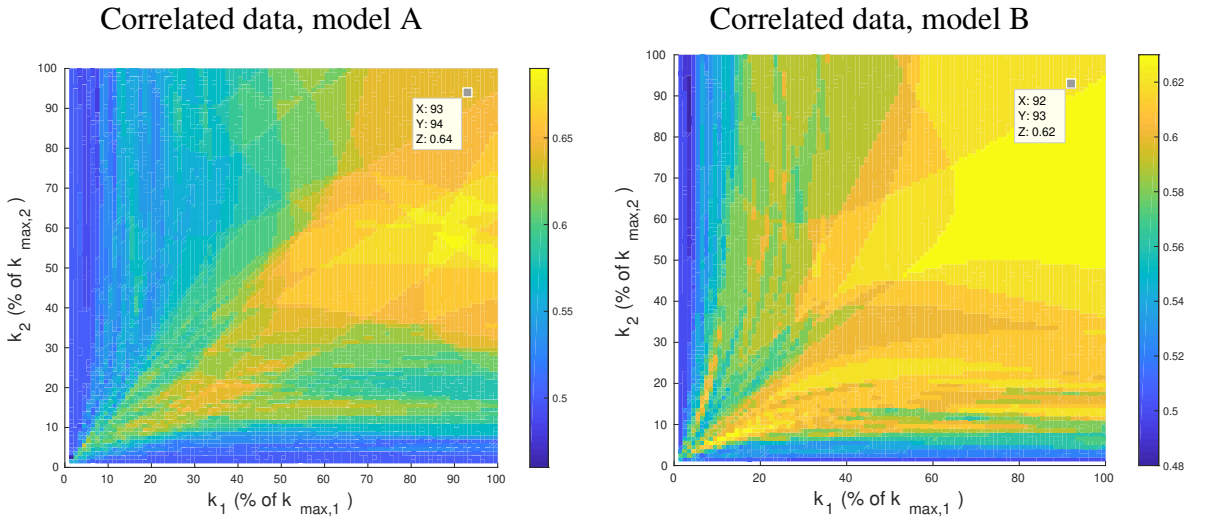

Fig. 6.4 LOOCV classification accuracy in (k_1, k_2) space for correlated synthetic data, with class means $\mu_1 = (0,0,\ldots,0)$ and $\mu_2 = (2.5,0,\ldots,0)$, population covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ of symmetric Toeplitz form defined by its first row $(p, p-1,\ldots,2,1)$, and covariate dimension $p = 50$. The hyperparameters for models A (left) and B (right) were determined via (6.42a),(6.42d) and labelled on the plot. The results are based on a single data realisation.

6.4.2 Overfitting

Next we illustrate the degree of overfitting for models A and B for correlated and uncorrelated synthetic datasets. Distributions described in [54, 128] were sampled to allow for comparison. A full description of the statistical features of these synthetic datasets is detailed in Table 6.3 for case 1 (uncorrelated) and case 8 (correlated case). We chose $C = 3$ classes of $N_z = 13$ samples each, for a broad range of data dimensions p.

Measuring training and validation classification performance via LOOCV on these data resulted in Figures 6.5 and 6.6, where each data-point is an average over 250 simulation experiments. The degree of divergence between the training and generalization curves is a direct measure of the degree of overfitting. We observe, in these figures, that model B overfits less for uncorrelated data, and model A overfits less for correlated data. This pattern is also seen more generally in Table 6.4, for a broader range of synthetic datasets. The uncorrelated results in Table 6.2 are too close to draw a conclusion and since they are for a single dataset, no error bars were generated. However, we note that all models still perform significantly above the random guess level on unseen data, even when $p \gg N_z$. For instance, for $p = 150$ (corresponding to $p/N_z \approx 11.5$) the Bayesian models can still classify some 80% of the unseen data correctly. We note the small size of the samples emphasizes we are not in the thermodynamic limit characteristic of statistical physics in Part I.

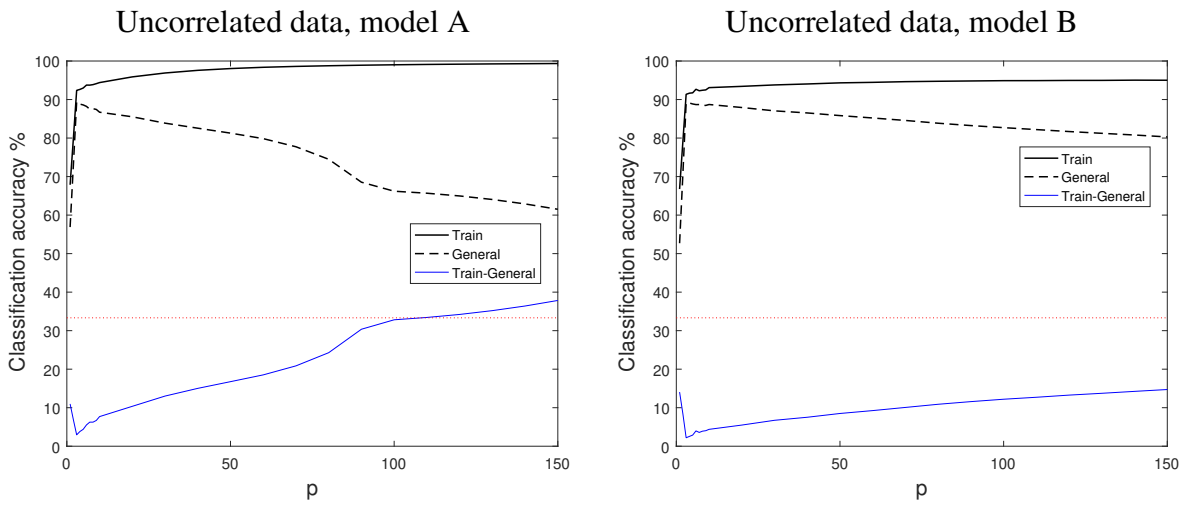

Fig. 6.5 Overfitting in models A (left) and B (right) as measured via LOOCV for uncorrelated data (case 1 in Table 6.3). $N_z = 13$ for each of the three classes. Absolute classification accuracies for training and generalization samples (black) and their difference (blue) is shown. The horizontal dotted line shows the baseline performance of a random guess classifier.

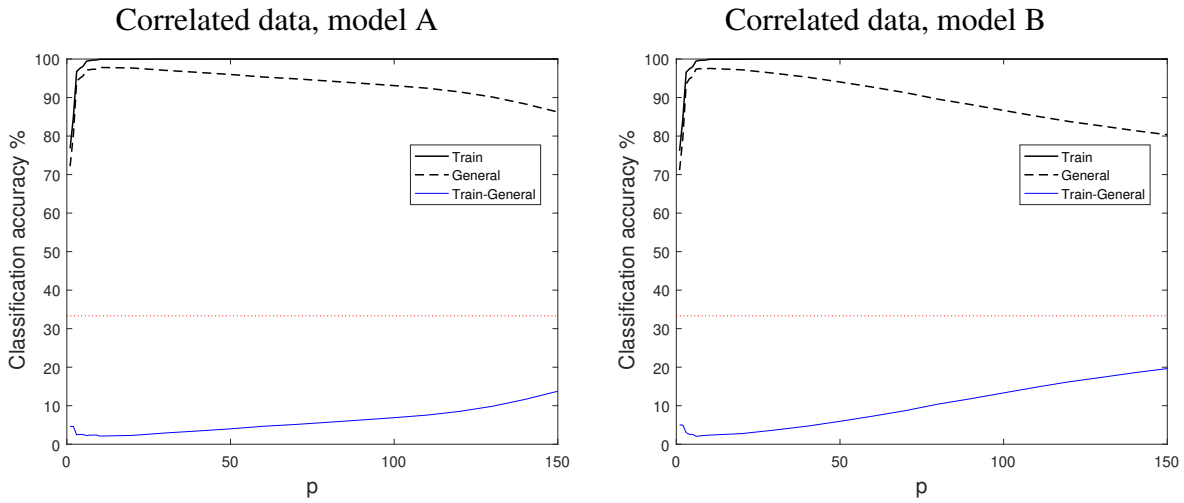

Fig. 6.6 Overfitting in models A (left) and B (right) as measured via LOOCV for correlated data (case 8 in Table 6.3). $N_z = 13$ for each of the three classes. Absolute classification accuracies for training and generalization samples (black) and their difference (blue) is shown. The horizontal dotted line shows the baseline performance of a random guess classifier.

6.5 Numerical results

We now investigate our classifier performance in greater detail using both synthetic and real data. The classification accuracy of models A and B, with hyperparameters optimized by evidence maximization, is compared to other successful state-of-the-art generative classifiers from [128]. These include the distribution-based Bayesian classifier (BDA7), the Quadratic Bayes (QB) classifier [16], and a non-Bayesian method, the so-called eigenvalue decomposition discriminant analysis (EDDA) as described in [11]. All three use cross-validation for model selection and hyperparameter estimation. The classifiers (our models A and B and the three benchmark methods from [128]) are all tested on the same synthetic and real datasets, and following the definitions and protocols described in [128], for a fair comparison. Model A differs from Quadratic Bayes [16] only in that our hyperparameters have been estimated using evidence maximization, as described in Section 6.3, rather than via cross-validation and is seen in Table 6.4 to have lower error rates than Quadratic Bayes in the majority of the synthetic datasets. In contrast, model B is mathematically different from both model A and Quadratic Bayes.

6.5.1 Implementation

The classifier was implemented in MATLAB [3]. The leave-one-out cross-validation pseudo-code is displayed in Algorithm 2. The rate limiting step in the algorithm is calculation of sample eigenvalues which scales with p^3. We note as the data dimension increases above 30,000, RAM storage considerations become an issue on typical PCs. For reproducibility, the MATLAB random number seed was set to *rng(1)* for synthetic data runs.

Algorithm 2 LOOCV classification

procedure LOOCV
  $\mathbf{X} \leftarrow$ design matrix, $\mathbf{y} \leftarrow$ class labels ▷ Import or generate data, $\mathbf{X} \in \mathbb{R}^{\mathbf{N}\times\mathbf{d}}$
  specify model A or B ▷ User input
  loop i = 1:N ▷ N = number of samples/rows
    $\mathbf{x}_0 = \mathbf{X}(\mathrm{i},:) \Rightarrow \mathscr{D} = \mathbf{X} \setminus \mathbf{x_0}$ ▷ $\mathscr{D}$ =training set
    calculate sample covariance matrix and eigenvalues from $\mathscr{D}$
    calculate hyperparameters ▷ (6.42b) (6.42d)
    predict $\leftarrow \mathbf{x}_0$ class prediction ▷ (6.57)
  end loop
  Create confusion matrix from **y** and **predict**
end procedure

[3] MATLAB 8.0, The MathWorks, Inc., Natick, Massachusetts, United States.

6.5.2 Synthetic data

We define datasets and precise testing protocols in order to test the performance of our Bayesian classifiers. By using the relevant definitions from [128], which used a set of ten synthetic datasets, all with Gaussian multivariate covariate distributions, we allow for a fair comparison across a range of classifiers.

Description of datasets - Cases 1-6. These represent uncorrelated data with choices for class-specific means and covariance matrices are detailed in Table 6.3. In the present study we generated data with exactly the same statistical features. Cases 1-6 were also used in the earlier work of [54].

Description of datasets - Cases 7-10. The remaining four cases represent correlated data and their covariance matrices are defined in terms of auxiliary random $p \times p$ matrices $\mathbf{R}_z$, with i.i.d. entries sampled from the uniform distribution on the interval $[0,1]$, according to either $\Sigma_z = \mathbf{R}_z^T \mathbf{R}_z$ or $\Sigma_z = \mathbf{R}_z^T \mathbf{R}_z \mathbf{R}_z^T \mathbf{R}_z$. By writing $\mathbf{R}_z = \mathbf{U}^T \text{ Diag}(\lambda_1, \ldots, \lambda_p)\ \mathbf{U}$ where $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbb{I}_p$ and $\{\lambda_i\}_{i=1}^p \in \mathbb{C}$, we find both forms of Σ_z are positive semi-definite matrices by construction. The difference is the latter form has a larger proportion of eigenvalues closer to the dominant one. The remainder are close to zero.

Datasets 7 and 9 have all class means at the origin, whereas each element of the class mean vectors from datasets 8 and 10 are independently sampled from a standard normal distribution.

Description of protocol. Each dataset has $C = 3$ outcome classes, and is separated into a training set, with $N_z = 13$ samples in each class, and a validation set, with $N_z = 33$ samples in each class. In terms of the balance N_z/p, this allows for a direct comparison with the dimensions used in [128].

Description of results. The results shown in Table 6.4 are all averages over 100 synthetic data runs with $p \in \{10, 50, 100\}$. Since all these synthetic datasets involve multivariate Gaussian covariate distributions, there is no model mismatch with any of the models being compared. Table 6.4 shows the classification error rates, as percentages of misclassified samples over the validation set. The variability of these for results for the models BDA7, QB and EDDA, i.e. the error bars in the classification scores, is not reported in [128] (where only the best classifier was determined using a signed ranked test). For completeness, we have included in this study the standard deviation of the error rate over the 100 synthetic data runs for our models A and B. Given that all experiments involved the same dimensions of datasets and similar average error rates, the error bars for the [128] results are expected to be similar to those of models A and B. We conclude from Table 6.4 that our models A and B perform

	Σ_1	Σ_2	Σ_3	μ_1	μ_2	μ_3
Case 1	$\mathbb{I}_p$	$\mathbb{I}_p$	$\mathbb{I}_p$	$(0,0,\ldots,0)$	$(3,0,\ldots,0,0)$	$(0,0,\ldots,0,3)$
Case 2	$\mathbb{I}_p$	$2\mathbb{I}_p$	$3\mathbb{I}_p$	$(0,0,\ldots,0)$	$(3,0,\ldots,0,0)$	$(0,0,\ldots,0,4)$
Case 3	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$(0,0,\ldots,0)$	$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\left(\frac{p-i}{\frac{p}{2}-1}\right)$	$\mu_{3i} = (-1)^i\mu_{2i}$
Case 4	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$(0,0,\ldots,0)$	$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\left(\frac{i-1}{\frac{p}{2}-1}\right)$	$\mu_{3i} = (-1)^i\mu_{2i}$
Case 5	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(p-i)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-(\frac{p-1}{2}))}{p-1}+1\right)^2$	$(0,0,\ldots,0)$	$(0,0,\ldots,0)$	$(0,0,\ldots,0)$
Case 6	$\Sigma_{ii} = \left(\frac{9(i-1)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(p-i)}{p-1}+1\right)^2$	$\Sigma_{ii} = \left(\frac{9(i-(\frac{p-1}{2}))}{p-1}+1\right)^2$	$(0,0,\ldots,0)$	$(\frac{14}{\sqrt{p}},\ldots,\frac{14}{\sqrt{p}})$	$\mu_{3i} = (-1)^i\mu_{2i}$
Case 7	$\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3$	$(0,0,\ldots,0)$	$(0,0,\ldots,0,0)$	$(0,0,\ldots,0,0)$
Case 8	$\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3$	$\mathrm{N}_p(0,1)$	$\mathrm{N}_p(0,1)$	$\mathrm{N}_p(0,1)$
Case 9	$\mathbf{R}_1^T\mathbf{R}_1\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3\mathbf{R}_3^T\mathbf{R}_3$	$(0,0,\ldots,0)$	$(0,0,\ldots,0,0)$	$(0,0,\ldots,0,0)$
Case 10	$\mathbf{R}_1^T\mathbf{R}_1\mathbf{R}_1^T\mathbf{R}_1$	$\mathbf{R}_2^T\mathbf{R}_2\mathbf{R}_2^T\mathbf{R}_2$	$\mathbf{R}_3^T\mathbf{R}_3\mathbf{R}_3^T\mathbf{R}_3$	$\mathrm{N}_p(0,1)$	$\mathrm{N}_p(0,1)$	$\mathrm{N}_p(0,1)$

Table 6.3 Description of synthetic datasets. The above class-specific means and covariance matrices were used to sample from a multivariate Gaussian distribution. These are the same data characteristics as those used in [128], reflecting varying degrees of correlation between variables.

Error rate (%)	*p*	*BDA7*	*QB*	*EDDA*	*model A*	*model B*
Case 1	10	13.2	19.2	11.2	12.0 ± 3.2	11.0 ± 2.8
	50	27.9	33.3	21.7	19.9 ± 4.6	15.6 ± 3.4
	100	35.8	31.1	24.8	32.6 ± 6.0	19.9 ± 4.3
Case 2	10	21.3	27.4	16.1	11.9 ± 3.4	11.4 ± 3.6
	50	26.8	42.6	12.5	9.3 ± 3.2	5.8 ± 2.2
	100	20.8	41.9	9.0	26.5 ± 5.6	3.6 ± 2.1
Case 3	10	10.4	35.0	9.1	27.2 ± 4.9	27.2 ± 5.5
	50	27.2	55.7	21.2	48.6 ± 5.0	49.2 ± 5.2
	100	46.9	56.4	27.7	55.4 ± 5.2	55.1 ± 4.9
Case 4	10	12.6	32.8	11.6	11.3 ± 3.5	11.1 ± 4.1
	50	22.5	30.9	17.0	22.5 ± 4.4	17.8 ± 4.0
	100	37.6	32.1	21.1	30.8 ± 5.2	21.9 ± 4.3
Case 5	10	4.1	15.0	4.4	12.8 ± 4.1	12.8 ± 3.5
	50	1.2	30.6	0.0	9.2 ± 3.4	5.6 ± 2.7
	100	0.2	38.3	0.1	10.9 ± 3.8	5.4 ± 3.4
Case 6	10	5.2	7.9	1.7	4.6 ± 2.3	4.4 ± 2.3
	50	0.5	26.5	0.0	3.9 ± 2.3	3.5 ± 2.4
	100	0.1	29.4	0.0	4.8 ± 2.5	4.5 ± 2.6
Case 7	10	19.5	22.8	19.7	20.0 ± 6.0	27.3 ± 7.4
	50	34.7	30.9	63.9	30.2 ± 5.0	44.7 ± 7.8
	100	40.0	35.2	64.8	35.2 ± 5.1	51.7 ± 7.8
Case 8	10	3.7	2.7	5.1	1.6 ± 1.9	1.5 ± 1.5
	50	9.2	3.5	25.5	4.4 ± 3.2	9.5 ± 5.0
	100	17.3	8.1	55.2	8.7 ± 4.4	23.9 ± 9.0
Case 9	10	1.5	0.9	1.0	0.9 ± 1.1	5.4 ± 6.8
	50	1.3	0.9	32.5	1.3 ± 1.2	16.9 ± 14.6
	100	2.9	2.8	67.0	1.5 ± 1.5	22.4 ± 15.3
Case 10	10	0.4	0.1	3.4	0.1 ± 0.6	0.2 ± 0.6
	50	1.7	0.9	32.4	0.8 ± 1.0	15.9 ± 13.6
	100	2.2	2.4	64.0	1.4 ± 1.2	23.4 ± 16.0

Table 6.4 Classification performance for synthetic datasets. Three generative Bayesian models, BDA7, QB and EDDA (results taken from [128]) are used as comparison with our models A and B. Error rates are the percentages of misclassified samples from the test dataset. The error bars for models A and B represent one standard deviation in the error rates, calculated over the 100 data realisations.

on average quite similarly to the benchmark classifiers BDA7, QB and EDDA. On some datasets model A and/or B outperform the benchmarks, on others they are outperformed. However, models A and B achieve this competitive level of classification accuracy without cross-validation, i.e. at a much lower computational cost.

Finally we determined in Section 6.3.3 that model B would outperform model A when there was no model mismatch. This is indeed the case for uncorrelated data (see cases 1-6 in Table 6.4). We have not yet determined why this does not hold for the correlated synthetic datasets of cases 7-10.

6.5.3 Real data

Description of datasets. Next we test the classification accuracy of our models against real datasets. We again chose the same datasets used in [128] to allow for direct comparison. They are publicly available from the UCI machine learning repository[4]. These are more testing than the synthetic data due to model mismatch i.e. data is unlikely to be generated from a multivariate Gaussian distribution. Three datasets were left out due to problems with matching the formats: *Image segmentation* (different number of samples than [128]), *Cover type* (different format of training/validation/test), and *Pen digits* (different format of training/validation/test). Before classification, we looked for identifying characteristics which could allow for retrospective interpretation of the results, e.g. occurrence of discrete covariate values, covariance matrix entropies, or class imbalances. None were found to be informative. No scaling or pre-processing was done to the data before classification.

Description of protocols. We duplicated exactly the protocol of [128], whereby only a randomly chosen 5% or 10% of the samples from each class of each dataset are used for training, leaving the bulk of the data (95% or 90%) to serve as validation (or test) set. The resulting small training sample sizes lead to $N_z \ll p$ for a number of datasets, providing a rigorous test for classifiers in overfitting-prone conditions. For example, the set *Ionosphere*, with $p = 34$, has original class sizes of 225 and 126 samples leading in the 5% training scenario to training sets with $N_1 = 12$ and $N_2 = 7$. We have used the convention of rounding up any non-integer number of training samples (rounding down the number of samples had only a minimal effect on most error rates). The *baseline* column gives the classification error that would be obtained if the majority class is predicted every time.

Description of results. We conclude from the classification results shown in Tables 6.5 and 6.6 (which are to be interpreted as having non-negligible error bars), that also for the real

[4]http://archive.ics.uci.edu/ml/index.php. Last accessed 21th October 2019

Error (%)	*N*	*class size*	*p*	*baseline*	*BDA7*	*QB*	*EDDA*	*model A*	*model B*
Heart	270	150,120	10	44.4	27.4	32.0	28.3	30.3	30.1
Ionosphere	351	225, 126	34	35.9	12.5	11.1	23.3	8.3	7.5
Iris	150	50,50,50	4	66.6	6.2	5.9	7.4	7.5	6.6
Pima	768	500,268	8	34.9	28.4	29.7	29.0	28.8	28.9
Sonar	208	97,111	60	46.6	31.2	33.7	34.8	34.9	33.8
Thyroid	215	150,35,30	5	30.2	7.9	9.1	8.6	7.6	7.9
Wine	178	59,71,48	13	60.1	7.9	16.9	8.2	15.6	16.0

Table 6.5 Average error rate using randomly selected 10% of training samples in each class. The remaining 90% of samples were used as a validation set. Error rates are the percentage of misclassified samples over this validation set.

Error (%)	*N*	*class size*	*p*	*baseline*	*BDA7*	*QB*	*EDDA*	*model A*	*model B*
Heart	270	150,120	10	44.4	30.6	38.5	33.9	38.8	39.6
Ionosphere	351	225, 126	34	35.9	16.9	16.1	26.0	10.3	8.8
Iris	150	50,50,50	4	66.6	6.9	7.6	9.40	12.8	11.4
Pima	768	500,268	8	34.9	29.7	32.7	30.7	30.3	30.8
Sonar	208	97,111	60	46.6	36.8	40.4	39.8	45.6	39.0
Thyroid	215	150,35,30	5	30.2	11.7	14.8	14.7	34.5	14.6
Wine	178	59,71,48	13	60.1	9.6	33.1	11.2	54.4	33.0

Table 6.6 Average error rate using randomly selected 5% of training samples in each class. The remaining 95% of samples were used as a validation set. Error rates are the percentage of misclassified samples over this validation set.

data, models A and B are competitive with the other Bayesian classifiers. The exceptions are *Ionosphere* (where models A and B outperform the benchmark methods, in both tables) and the datasets *Thyroid* and *Wine* (where in Table 6.6 our model A is being outperformed). Note that in Table 6.6, *Thyroid* and *Wine* have only 2 or 3 data samples in some classes of the training set. This results in nearly degenerate class-specific covariance matrices, which hampers the optimization of hyperparameters via evidence maximization. Model B behaves well even in these tricky cases, presumably due to the impact of its additional hyperparameter $\gamma_{0z} = p/\hat{\mathbf{x}}_z^2$. As expected, upon testing classification performance using leave-one-out cross-validation (details not shown here) rather than the 5% or 10% training set methods above, all error rates are significantly lower.

Examining the results from Sections 6.5.2 and 6.5.3 does not lead us to conclusions on when one specific model outperforms the other. Unfortunately the theoretical finding of the outperformance of model B in Section 6.3.3 does not help us since these datasets were not generated from a Gaussian distribution. A possible empirical approach to understanding the conditions where models perform well is generate synthetic data from a multivariate t-distribution. The degrees of freedom, ν, provides a parametrization of model mismatch with the limiting case of $\nu \to \infty$ recovering the Gaussian distribution.

6.6 Discussion

In this chapter, we considered generative models for Bayesian classification in high-dimensional spaces. Our aim was to derive expressions for the optimal hyperparameters and predictive probabilities in closed form. Since the dominant cause of overfitting in the classification of high-dimensional data is using point estimates for high-dimensional parameter vectors, we believe that by carefully choosing Bayesian models for which parameter integrals are analytically tractable, we will need point estimates only at hyperparameter level, reducing overfitting.

We showed that the standard priors of Bayesian classifiers that are based on class-specific multivariate Gaussian covariate distributions can be generalized, from which we derive two special model cases (A and B) for which predictive probabilities can be derived analytically in fully explicit form. Model A is known in the literature as Quadratic Bayes [16], whereas model B is novel and has not yet appeared in the literature. In contrast to common practice for most Bayesian classifiers, we use evidence maximization [91] to find analytical expressions for our hyperparameters in both models. This allows us to find their optimal values without needing to resort to computationally expensive cross-validation protocols.

We found that the alternative (but significantly faster) hyperparameter determination by evidence maximization leads to hyperparameters that are generally very similar to those obtained via cross-validation, and that the classification performance of our models A and B degrades only gracefully in the 'dangerous' regime $N \ll p$ where we would expect extreme overfitting. We compared the classification performance of our models on the extensive synthetic and real datasets that have been used earlier as performance benchmarks in [127, 128]. Interestingly, the performance of our models A and B turned out to be competitive with state-of-the-art Bayesian models that use cross-validation, despite the large reduction in computational expense. This will enable users in practice to classify high-dimensional datasets quicker, without compromising on accuracy.

This work suggests that the analytical approach merits further investigation. Calculating the predictive probability for arbitrary γ_{0z}, γ_{1z} values remains to be done. The main obstacle being the resulting symbolic integration. We believe this could lead to interesting analytically tractable classification models.

Implementing methods to mitigate the effect of class imbalance is prohibitively expensive for large p. Using the same approach but with a discriminative model (see [122] for the uncorrelated case) would avoid the need for including $p \times p$ covariance matrices. This work is in early stages and not included in this thesis.

Finally our implementation of the Bayesian classifier involves evaluating all terms in the predictive probability expression (6.57) directly including eigenvalues of the $p \times p$ sample covariance matrices and p-dimensional quadratic forms. It may be possible to replace exact calculations with numerical approximations [98] without sacrificing classification accuracy.

Chapter 7

Improved resection margins in breast-conserving surgery using Terahertz Pulsed imaging data

Having shown our multivariate Bayesian classifier to be competitive on synthetic and sanitized data, we put forward a more rigorous test with data (with $p \sim N$) from a recent study. Before describing the methodology, we motivate the medical problem and give some background to the current work.

7.1 Introduction

Breast cancer is the most frequent cancer among women worldwide [52] and the leading cause of cancer death in females, accounting for 23% of the total cancer cases and 14% of the cancer deaths [77]. Surgery to remove the primary tumour is one of the main curative treatments in breast cancer which has evolved over time from radical surgery to more conservative approaches, such as breast conserving surgery [53, 116].

Good surgical outcomes are related to the accuracy in the resection of the tumour margins (edges of tissue sample) during surgery. Currently, a significant number of patients need to undergo additional surgery to obtain clear margins and/or remove remaining lymph nodes in the axilla, causing a significant physical and psychological morbidity in patients [50]. To date, histopathology, which occurs after surgery, is the gold standard to determine tumour margins. Hence, new techniques are being investigated to assess tumour margins intraoperatively. Terahertz (THz) radiation is one such tool (see Figure 7.1); it does not produce harmful

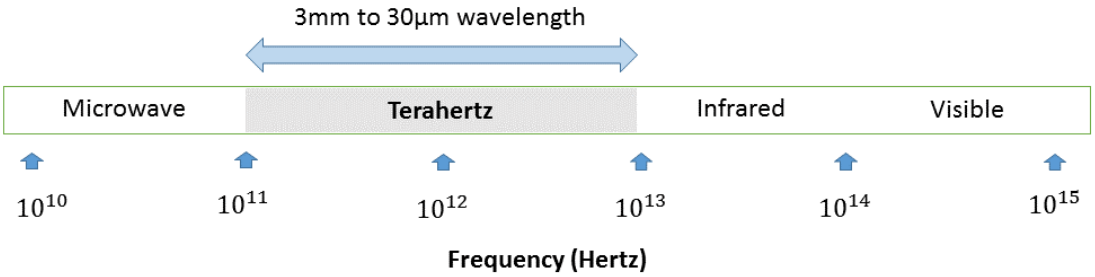

Fig. 7.1 Electromagnetic spectrum showing frequency of Terahertz radiation.

ionization in biological tissues and therefore has been applied in cancer studies previously [51].

In order to implement this technology to assess tumour margins and sentinel lymph nodes in breast cancer patients intraoperatively, Teraview Ltd. (Cambridge, UK) developed a handheld THz Probe to be used in breast-conserving surgery, eventually aiming to reduce the number of reoperations. In [51], a feasibility study was performed to determine the capability of the handheld THz Probe to distinguish between different breast cancer tissues ex vivo. The THz raw data obtained was complex, multidimensional and noisy. There is a lack of standardised procedures for data acquisition, post processing, image analysis and interpretation of the data generated with this technology. Therefore, [51] used two different statistical approaches to discriminate between different breast tissue samples based on Terahertz pulsed imaging (TPI) data produced by the Teraview probe (see Figure 7.3).

Using the same dataset as [51], we aim to improve the prediction capabilities of the technology by utilising the novel Bayesian classifier from the previous chapter to discriminate between tumour, fibrous and adipose tissue. We first follow the protocol from [51] to allow for direct comparison of classification methods with the same datasets. Next we adjust our method in line with the relevant clinical objective. Both sets of results are outlined in Section 7.4 for transparency.

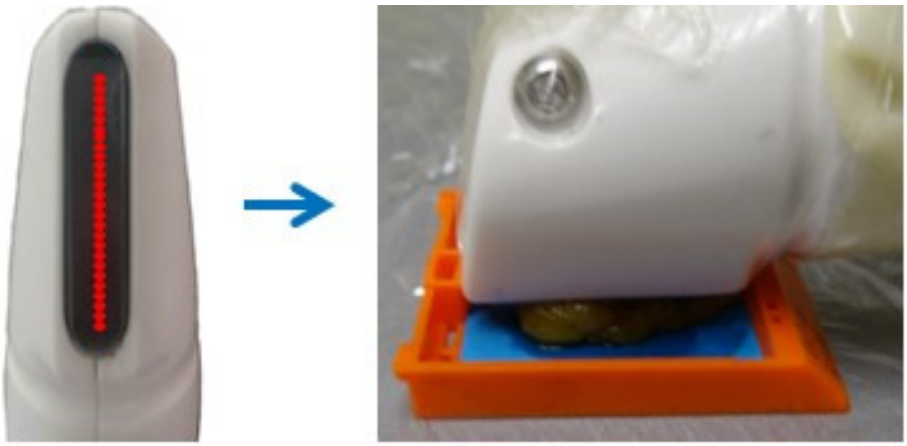

Fig. 7.2 Schematic description of the raw data acquisition. TPI handheld probe measurement of tissue sample positioned in histology cassette. Residual THz pulses are received by each pixel from the tissue producing typical TPI waveforms per pixel. Images are courtesy of the study (REC12-EE-0493) and with kind permission from TeraView Ltd., Cambridge, CB4 0WS, UK.

7.2 Methods

7.2.1 Technology

The handheld Terahertz pulsed imaging probe device (Teraview Ltd - see Figure 7.2) has 26 pixels. Each pixel transmits Terahertz electromagnetic radiation of frequency 0.1-2.0 THz to the biological sample and records any residual signal transmitted from the tissue. Technical details can be found in [64].

7.2.2 Data acquisition

Imaging data was produced from scanning 46 freshly excised breast cancer samples obtained from 30 breast cancer patients treated at Guy's and St Thomas' NHS Foundation Trust (GSTT) following a breast conservation surgery or mastectomy between August 2013 and August 2014. This dataset was acquired as part of a feasibility study to test the utility of the handheld TPI probe to discriminate between breast tissue types (REC 12-EE-0493).

Labelling of samples. The breast surgical specimens were sampled, collecting areas with high lesions into histology cassettes without compromising patient's clinical assessment after surgery. Each sample was assessed for detailed histopathology by two experts blinded for the patient details that labelled the percentage of adipose, fibrous and tumour at pixel level.

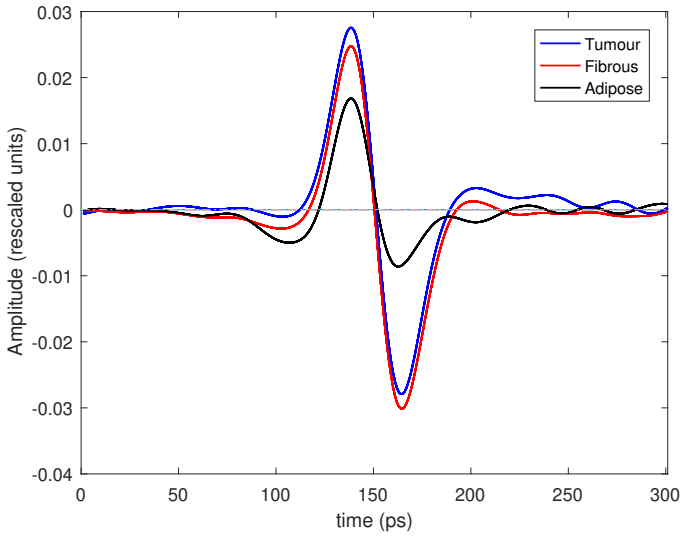

Fig. 7.3 Terahertz Pulsed Imaging waveform for tumour, fibrous and adipose cells. For each tissue type present in the sample, the TPI waveform presented a different shape: tumour (blue), fibrous (red) and adipose cells (black).

These were later converted into class labels suitable for statistical classification (see Section 7.3).

Conversion to digital form. The fresh breast tissue samples were then scanned with the probe (see Figure 7.2) in the pathology suite using air as a reference [51]. Each of the 26 pixels of the probe transmitted many pulses to the tissue sample and received the residual data. The multiple waveforms for each pixel were averaged within the device to produce a single data sample $\mathbf{x_i}$. Put another way, the arithmetic average of the amplitude measurements was calculated over many waveforms for each time point. Combining these for all time points results in a single waveform or data sample $\mathbf{x_i} \in \mathbb{R}^p$. If the raw unaveraged information from each pulse was preserved, a more accurate idea of the signal variability could have been obtained. The size and orientations of the samples were heterogeneous resulting in varying numbers of pixels per tissue.

Form of data used for classification. The waveform produced by each pixel had 674 data points. Removal of the outermost points, which were either zero or noise [64], resulted in 301 data points per waveform (see Figure 7.3). The final dataset consisted of predictors, $\{\mathbf{x_i}\}_{i=1}^{257}$ where $\mathbf{x_i} \in \mathbb{R}^{301}$ and associated response variables $y_i \in \{1,2,3\}$ corresponding to tumour (115 samples), fibrous (100 samples) and adipose (42 samples).

7.2.3 Data pre-processing

Our aim is to discriminate each impulse function (waveform) per pixel per sample into tumour, fibrous or benign breast tissue. However it is difficult to do so visually (see Figure 7.3). Statistical analysis is therefore required.

For problems where the number of covariates are high, such as the current one, feature selection is often used to choose the most informative features and often results in more robust discrimination. In [64], the Support Vector Machine [14] methodology used heuristic methods to determine which features had the greatest effect on classification results. For the Naive Bayesian classifier, also used in [64], overfitting is not as large a concern. Gaussian deconvolution, common in image processing, was used before sending the data to the classifier. In this present chapter, we found that classification based on the raw waveform data produced the best results.

Throughout this paper, we distinguish between a tissue sample, which is a physical sample from a patient, and a data sample, understood in the statistical sense i.e. data from a single pixel of a single sample. In this study there were 46 tissues samples collected which produced 257 data samples. The low number of data samples available for analysis is due to the orientation and size of the tissue samples in the histopathological cassette. This average of pixels is consistent with the original paper. In other words, the number of pixels per tissue sample was less than six, resulting in a challenging problem for the classification method. Our decision to use a Bayesian classification method was driven by the low ratio of samples to covariates.

7.3 Classification

The multivariate Bayesian classification algorithm of the previous chapter (and the associated peer-review article [124]) was applied to the TPI waveform data to discriminate benign from malignant breast tissue. The classifier used supervised learning techniques with the class labels derived from the histopathologically defined tissue content. The trained classifier was used to predict the classes of new observations.

In statistical learning, a classifier is a function which maps a data point to a class prediction. There are four stages to a classification problem. The first is model selection. We use the framework outlined in the previous chapter. The second stage is to train our chosen model on the training data. In our case, this involves estimating the hyperparameters for each class: tumour, fibrous and adipose. In the third stage, the classifier uses a probabilistic approach assuming that the observations in each class were generated by a multivariate Gaussian distribution specific to that class. Predictive probabilities, $p(y_0|\mathbf{x}_0, \mathscr{D})$, of a new

data sample belonging to a certain class are calculated. Lastly, decision theory is applied to translate those resulting probabilities into class predictions by choosing the class with the highest probability conditioned on the data sample. In all 257 cases, the classifier assigned a greater than 95% probability to the assigned class reflecting a high degree of certainty. While traditional statistical methods produce probability point estimates, probability distributions are produced in our Bayesian approach. This allows us to quantify the uncertainty in the predictions made.

The performance of the classifier was evaluated using leave-one-out cross-validation (LOOCV) i.e. training of the classifier by leaving out a single data sample, learning parameters from the remaining data samples and then using these estimated parameters to classify the data sample that was left out. This process was repeated for all the samples. The results were compiled to estimate accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) to distinguish malignant from benign tissue. These terms are defined via the confusion matrix in Appendix A.7.

This approach is a standard statistical inference procedure for small datasets i.e. those without separate training and validation sets. However it is recognised that including data samples from the same patient is not clinically realistic and will likely flatter the classification accuracy metrics. Therefore we repeat the cross-validation process but leaving out whole tissue samples rather than just data samples. This is closer to practical clinical applications but leads to lower predictive accuracy.

Typically, data with more covariates, p, than samples, N, leads to overfitting [26]. This is the case with our data where $N = 257$ and $p = 301$. This is somewhat mitigated by our fully Bayesian classification approach. Inference methods such as maximum likelihood work well when $N \gg p$ but are prone to overfitting when $N \leq p$. The classifier used in this chapter avoids taking point estimates of unknown parameters. Instead it treats them in a Bayesian way by integrating over them. By carefully choosing the prior probability distributions, a closed-form for the predictive probabilities is derived. The final estimation of hyperparameters is done analytically. An added advantage is that computationally expensive cross-validation methods are avoided allowing us to handle larger datasets.

The Terahertz Pulsed Imaging data could be categorised under three different scenarios depending on the particular research question assessed. The class label for each sample will be determined by:

- Scenario 1: pixels were marked as tumour when containing any amount of cancer cells (if the tumour percentage was greater than zero, the TPI impulse function was considered as tumour, irrespective of the content of fibrous or adipose tissue), otherwise

the tissue content of the pixel was defined by the highest percentage of fibrous or adipose tissue;

- Scenario 2: pixels were marked as tumour when containing any amount of cancer cells (if the tumour percentage was greater than zero, the TPI function was considered as tumour, independently of the content of fibrous or adipose tissue), otherwise tissue was considered to be benign. In this scenario, adipose and fibrous tissues were grouped together;

- Scenario 3: pixels were marked based on their tissue content, so the tissue content of each pixel was defined by the highest percentage.

The project goal was to discriminate malignant tissue from benign breast tissue suggesting a choice of scenarios 1 or 2. The fibrous and adipose waveforms are clearly distinct (see Figure 7.3). To allow our classifier to pick up these differing statistical characteristics, scenario 1 was the preferred option in the study. This results in class sizes of 115, 100 and 42 (totalling 257 data samples). Put another way, grouping fibrous and adipose samples together, as in scenario 2, would lose distinguishing features between them.

The original THz time domain pulses were assessed in their ability to discriminate tumour from healthy (fibrous/adipose) breast tissue using our probabilistic Bayesian classification algorithm.

As we have seen in Part I, classification methods with imbalanced class sizes tend to favour the majority class [141]. This is a well known phenomenon in classification literature. The problem frequently occurs in medical datasets where the rare or diseased cases have many fewer samples. In our case, the data collected focused on tumorous tissue and hence the adipose class has many fewer samples.

Methods to compensate for this effect can be grouped into pre-processing methods or algorithmic changes. Pre-processing can be achieved by over-sampling the minority class [23] or under-sampling the majority class [38]. On the other hand, the classification algorithm itself can be changed [141]. An example is to specify a cost function for misclassification to mitigate the class imbalance effect. We have not attempted these methods in this chapter.

7.4 Results

Classification accuracy is first displayed for the three class analysis (tumour, fibrous and adipose) in Table 7.1. A common format for displaying classification results is the confusion matrix which is introduced in Appendix A.7. Three classes were used so the classifier could

more accurately identify characteristic fingerprints of tumour, fibrous and adipose waveforms. The number of correctly predicted samples was 218 out of 257 which gave a classification accuracy of 85%.

	Predict 1	Predict 2	Predict 3
True class 1	110	5	0
True class 2	5	95	0
True class 3	2	27	13

Table 7.1 Confusion matrix for the three class analysis using data samples. The classification results are split into three classes: class 1 represents tumour, class 2 fibrous and class 3 adipose tissue.

However the clinically important decision was whether the sample contains tumour or not. Classes 2 and 3 represented benign tissue. The fibrous and adipose classes were collapsed into a "non-tumour" class by simply adding relevant entries in the confusion matrix to obtain Table 7.2. This resulted in classification accuracy, sensitivity and specificity of 95%, 96% and 95% respectively. In addition the PPV and NPV values were 94% and 96%.

	Predict 1	Predict 2 & 3
True class 1	110	5
True class 2 & 3	7	135

Table 7.2 Aggregated confusion matrix using data samples. Classes 2 and 3 representing fibrous and adipose tissue have been aggregated into one class in line with the clinical objective.

The methodology used so far was purposefully selected to match the previous study of [64]. Therefore we include a comparison between it and our Tables 7.1 and 7.2.

Classifier	Bayesian method	Support Vector Machine [64]	Naive Bayes [64]
Accuracy	**95%**	75%	69%
Sensitivity	**96%**	86%	89%
Specificity	**95%**	66%	53%
PPV	**94%**	67%	60%
NPV	**96%**	85%	86%

Table 7.3 Comparison of classification results. The Bayesian method column is the result of our classifier from Chapter 6. Models A and B produced the same results in this case. The Support Vector Machine and Naive Bayes columns are results from the previous study on the same dataset [64]. Figures in bold are the best for that metric.

Having shown this comparison, we proceed with the more clinically relevant results for cross-validation by *tissue* sample (Tables 7.4 and 7.5). The class imbalance problem can clearly be seen for the minority adipose samples. Since the classifier only has access to the current study data, it cannot adequately characterize the adipose data signal. This is exaggerated when a block of adipose data is removed in the cross-validation process. This effect mistakenly classifies adipose samples as fibrous which is not a clinically relevant problem. A straight-forward remedy would be to augment the data set with further tissue samples from healthy patients.

	Predict 1	Predict 2	Predict 3
True class 1	75	40	0
True class 2	38	62	0
True class 3	2	40	0

Table 7.4 Confusion matrix for the three class analysis using tissue samples. The classification results are split into three classes: class 1 represents tumour, class 2 fibrous and class 3 adipose tissue. Cross-validation by tissue sample.

	Predict 1	Predict 2 & 3
True class 1	75	40
True class 2 & 3	40	102

Table 7.5 Aggregated confusion matrix using tissue samples. Classes 2 and 3 representing fibrous and adipose tissue have been aggregated into one class in line with the clinical objective. Cross-validation by tissue sample.

Next we summarise the classification results on the same dataset using our method and those from previous studies [64] (see Table 7.3).

7.5 Comparison to existing methods

Different intraoperative margin assessment (IMA) tools are being investigated to assess tumour resection margins. Some of these techniques are clinically established, such as frozen section analysis [22, 78, 84], specimen radiography, intraoperative ultrasound, touch imprint cytology and optical spectroscopy. However these IMAs present diverse performance and limitations in terms of accuracy, speed, cost, and reliability. Therefore, new IMA tools are currently under development including Raman spectroscopy [66, 88], microcomputed CT, mass spectroscopy [130], radiofrequency spectroscopy [118] or fluorescence imaging.

	Number of patients	Accuracy	Sensitivity	Specificity
Frozen section analysis (FSA)	46-1327	**84-98 %**	78-91%	**92-98%**
Specimen radiography	12-119	33-84%	45-61%	77-89%
Intraoperative ultrasound	81-225	62-80%	36-79%	66-91%
Touch imprint cytology	27-510	78-99%	**71-97%**	**90-98%**
Optical spectroscopy	20-179	75-94%	74-91%	**65-96%**
Support Vector Machine	257	75%	86%	66%
Naive Bayesian method	257	69%	89%	53%
New Bayesian method	257	**95%**	**96%**	**95%**

Table 7.6 Comparison to existing methods. These results are taken from [129]. The number of patients/samples and accuracy are the ranges across all studies reviewed. The sensitivity and specificity are the result of their meta-analysis. Figures in bold are the best for that metric.

Our classification results were materially better than the performance of IMAs currently used, as reported by a recent meta-analysis [129] (see Table 7.6).

7.6 Discussion

From the previous section, our classifier seems to be learning most of the signal in the data. Its results are comparable to Frozen Section Analysis and it performs better on all measures than the previous classifiers from [51] on the given dataset. This previous analysis used Support Vector Machines combined with heuristics to select suitable features. Our method has the advantage of using all the available data and requiring no heuristic choices. Additionally, there are no ad hoc choices for estimating the hyperparameters since our approach is analytic.

The dataset used has class sizes of 115, 100 and 42. This class imbalance will lead to worse classification than a balanced dataset. We plan to use data pre-processing methods to help with class imbalance. Given this involves multiple runs with different samples removed, it requires much more time compared with no pre-processing.

7.6.1 Next steps

Within the tumour class label, there are actually three tumour subtypes (invasive ductal carcinoma, invasive lobular carcinoma and invasive tubular carcinoma) where each of these may have statistically different TPI waveforms. Our classifier is certainly able to discriminate more than three classes but the problem of too few data samples relative to the number of covariates leads to insufficient characterisation of each class. If sufficient numbers of each type were present in the data, it is likely that classification accuracy would improve.

The composition of human breast tissue can vary with patient age. Having access to additional data such as patient age is likely to improve classification accuracy. Our method of leave-one-out cross-validation is a standard statistical learning procedure given small datasets. It was used to allow comparison to existing literature, specifically [64]. However it is recognised that including data samples from the same patient is not clinically realistic. This likely flattered our classification accuracy metrics. This should be compensated by increasing the overall training size.

Our data was obtained from ex vivo tissue samples. Since Terahertz radiation is non-ionising, it would be interesting to re-run the analysis using data from in vivo clinical trials. Lastly, our analytical methods are sufficiently general to be applied to other tumour types.

Chapter 8

Conclusion

In this thesis, we investigate the three stages of the inference hierarchy: Maximum likelihood (ML), maximum a posteriori (MAP) and fully Bayesian inference. These move us progressively towards the ultimate goal of accurate statistical inference in high dimensions. At each stage, we critically examine accepted statistical practices and the conditions where they are no longer valid. The two main goals of inference are to estimate model parameters and to correctly predict the output for a new data sample. Here we take two different perspectives to these goals both of which focus on the high-dimensional data regime: statistical physics in Part I and Bayesian inference in Part II. Interesting behaviour is found across these methods depending on data characteristics such as ratio of p/N, the degree of correlation between features and the presence of imbalanced class sizes.

ML inference is an established method, known to be optimal in the classical statistical regime of $\zeta = p/N \to 0$. It has been a standard inference method for a century and is intuitively appealing. However, despite its advantages, we find, in Part I, that the inference of model parameters is systematically biased when considering three common generalized linear models (GLMs): linear, logistic and Cox regression. In particular, we find that the Maximum Likelihood procedure always predicts model parameters to be more extreme than their true counterparts resulting in a deterioration in prediction accuracy. We use methods of statistical physics to estimate the extent of this bias. Interpretation of our results leads to regions where the macroscopic parameters of our problem diverge. These correspond to regions where inference is either unreliable or no longer possible.

To explore data structures beyond the ζ regimes infeasible under ML, we introduce a prior probability on the model parameters which has the effect of regularizing the inference problem. Our choice of $L2$ regularization allows integrals over the regression coefficients to be evaluated analytically. In turn, this permits the derivation of a set of equations linking various macroscopic observables of the original inference problem. We discuss other regu-

larization choices such as $L0$ or $L1$ priors which may be more relevant for inference when parameters are sparse and also a justification for our current choice. The resulting MAP theory predicts the absence of a phase transition, characteristic of ML, and finds relationships between the following: bias of both the mean and variance of inferred regression outcomes; the ratio $\zeta = p/N$ and the amount of regularization η. Prediction of an unseen data sample using systematically exaggerated regression coefficients is clearly problematic and always acts to reduce the prediction accuracy. Regularization is shown to mitigate this effect in all three different GLMs considered. In addition, our formalism allows the investigation of the related problem of class imbalance by separating the intercept term from other regression coefficients. The common occurrence of biomedical data with imbalanced class-sizes makes this particularly relevant.

The work of Part I can be extended along both practical and theoretical paths. Our theory has been built upon the idealized scenario of knowledge of the underlying data-generating model. The effect of model mismatch is an important, particularly for real-world applications, and will be the subject of future work. Further, to put our work into practice, the variance of the true regression parameters, $\tilde{S}$, and the population covariance matrix, $\mathbf{A}$ need to be estimated. The former is available through the existing order parameters. The latter can be understood by noting that the dependence on $\mathbf{A}$ is only via integrals over its eigenvalue spectrum. Further work is planned to estimate these through either inverting a discretized Marčenko-Pastur equation [42] or by matching moments of the eigenvalue spectra [19].

In addition to practical considerations, there is a natural progression to our theoretical analysis. We can generalize our formalism by delaying the specification of the model until the order parameter equations have been derived. This avenue was recently explored in [27]. Reassuringly, the results match those of this thesis for specific model choices. For time-to-event models, analyzing censored data is still an open problem along with finding better solutions for base hazard rate and for non-Gaussian inner products $\beta \cdot \mathbf{z}$. This will allow our analysis to be applied to real data sets and is in progress. Further, there are a range of interesting problems to consider which have multiple linear predictors including multinomial regression [105], multiple risks in survival analysis [33] and multilayer neural networks [87]. This complicates the theory by requiring matrix order parameters where there are scalar ones in our current work (and tensors instead of the current matrices).

To be clear, MAP inference, which calculates the mode of the posterior distribution, is also not optimal when ζ is different from zero. We are therefore naturally led to consider the fully Bayesian approach in Part II by attempting to integrate the posterior distribution. Numerical methods are problematic since sampling from a high-dimensional posterior distribution does not scale well with data dimension. Instead, we proceed analytically by carefully choosing

a generalized form of the Wishart prior which still allows for symbolic integration of the posterior probability. The result is two closed form expressions for the predictive probability of a multi-class generative classifier. The first recovers existing work from the literature, the second is novel. As a by-product of our approach, we derived an algorithm to estimate optimal hyperparameters via evidence maximization. After extensive comparisons with synthetic and example datasets, we apply our Bayesian classifier to medical data from a recent breast cancer study. This is important since experimental data is very unlikely to fit our assumptions. Despite this likely model mismatch, we find our method outperforms previous analysis on this specific dataset. However, more work is needed to understand under what conditions and on which data sets our methods outperform.

Our work suggests numerous future research paths. Our Bayesian method can be extended by widening the family of prior distributions with alternative hyperparameter values but we have not made meaningful progress with the integrals produced under this generalization. Some hyperparameters choices were discarded due to excessive computational overhead. Finding efficient algorithms for the related optimization may lead to further improvement in classification accuracy. A different approach is to analyze discriminative classifiers where we define a parametrized form of $p(C|\mathbf{x})$ (rather than assumptions on $p(C,\mathbf{x})$ in the generative case) and consider a suitable range of analytically tractable priors.

The task facing scientists today is how to apply these inference methods to accurately draw conclusions from medical images, clinical trials and biomedical experiments. Unfortunately, the trend has been from orthodox statistical thinking, which relies on underlying models, to purely algorithmic approaches, which all too often rely on heuristics. The focus has been on finding patterns within the data without understanding where these methods break down. We hope this thesis has taken a step in the direction of understanding some of the obstacles facing scientists attempting to infer information from vast swathes of biomedical data.

References

[1] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann Stat*, 6(4):701–726.

[2] Ai-Jun, Y. and Xin-Yuan, S. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2):215–222.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE T Automat Contr*, 19(6):716–723.

[4] Anderson, J. and Richardson, S. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1):71–78.

[5] Anderson, T. (1984). Multivariate statistical analysis. *Wiley and Sons, New York, NY*.

[6] Bartlett, M. (1953a). Approximate confidence intervals. *Biometrika*, 40(1/2):12–19.

[7] Bartlett, M. (1953b). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, 40(3/4):306–317.

[8] Battey, H. and Cox, D. (2018). Large numbers of explanatory variables: a probabilistic assessment. *P Roy Soc A-Math Phy*, 474(2215):20170631.

[9] Baxter, R. (2016). *Exactly solved models in statistical mechanics*. Elsevier.

[10] Bender, R. et al. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat Med*, 24(11):1713–1723.

[11] Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J Am Stat Assoc*, 91(436):1743–1748.

[12] Berger, J., Bernardo, J., et al. (1992). On the development of reference priors. *Bayesian statistics*, 4(4):35–60.

[13] Bishop, C. (2001). *Pattern Recognition and Machine Learning*. Springer, New York.

[14] Boser, B. E. et al. (2003). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152.

[15] Breslow, N. (1972). Contribution to discussion of paper by DR Cox. *J Roy Statist Soc, Ser B*, 34:216–217.

[16] Brown, P. et al. (1999). Discrimination with many variables. *J Am Stat Assoc*, 94(448):1320–1329.

[17] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

[18] Bulso, N. et al. (2019). On the complexity of logistic regression models. *Neural Comput*, 31(8):1592–1623.

[19] Burda, Z. et al. (2005). Spectral moments of correlated Wishart matrices. *Phys Rev E*, 71(2):026111.

[20] Cartwright, D. and Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychol Rev*, 63(5):277.

[21] Cattaneo, A. et al. (2013). Candidate genes expression profile associated with antidepressants response in the GENDEP study: Differentiating between baseline predictors and longitudinal targets. *Neuropsychopharmacol*, 38(3):377–385.

[22] Cendán, J. et al. (2005). Accuracy of intraoperative frozen-section analysis of breast cancer lumpectomy-bed margins. *J Am Coll Surgeons*, 201(2):194–198.

[23] Chawla, N. et al. (2002). SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*, 16:321–357.

[24] Chopra, P. et al. (2010). Improving cancer classification accuracy using gene pairs. *PLoS One*, 5(12):e14305.

[25] Concato, J. et al. (1995). Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. *J Clin Epidemiol*, 48(12):1495–1501.

[26] Coolen, A. et al. (2017). Replica analysis of overfitting in regression models for time-to-event data. *J Phys A-Math Theor*, 50(37):375001.

[27] Coolen, A. et al. (2020). Replica analysis of overfitting in generalized linear models. *arXiv preprint arXiv:2004.06329*.

[28] Coolen, A. and Saad, D. (1999). Dynamics of supervised learning with restricted training sets. In *Adv Neur In*, pages 197–203.

[29] Copelli, M. and Caticha, N. (1999). Universal asymptotics in committee machines with tree architecture. In *On-line learning in neural networks*, pages 165–181. Cambridge University Press.

[30] Courvoisier, D. et al. (2011). Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *J Clin Epidemiol*, 64(9):993–1000.

[31] Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans Electron*, EC-14(3):326–334.

[32] Cox, D. (1972). Regression models and life-tables. *J Roy Stat Soc B Met*, 34(2):187–202.

[33] Cox, D. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

[34] Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J Roy Stat Soc B Met*, 49(1):1–18.

[35] Cox, D. and Snell, E. (1968). A general definition of residuals. *J Roy Stat Soc B Met*, 30(2):248–275.

[36] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numer Math*, 31(4):377–403.

[37] Dobson, A. and Barnett, A. (2008). *An introduction to generalized linear models*. Chapman and Hall/CRC.

[38] Drummond, C. et al. (2003). C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11. Citeseer Washington DC.

[39] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

[40] Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

[41] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J Am Stat Assoc*, 68(341):117–130.

[42] El Karoui, N. et al. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann Stat*, 36(6):2757–2790.

[43] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high-dimensional feature space. *Stat Sinica*, 20(1):101.

[44] Fan, Y. et al. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *J Mach Learn Res*, 20(77):1–33.

[45] Fan, Y. and Tang, C. (2013). Tuning parameter selection in high-dimensional penalized likelihood. *J Roy Stat Soc B*, 75(3):531–552.

[46] Feller, W. (1971). *An Introduction to Probability theory and its application Vol II*. John Wiley and Sons.

[47] Finak, G. et al. (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med*, 14(5):518–527.

[48] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

[49] Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philos T R Soc Lond*, 222(594-604):309–368.

[50] Fitzal, F. and Gnant, M. (2006). Breast conservation: Evolution of surgical strategies. *Breast J*, 12(s2):S165–S173.

[51] Fitzgerald, A. et al. (2006). Terahertz pulsed imaging of human breast tumors. *Radiology*, 239(2):533–540.

[52] Fitzmaurice, C. et al. (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study. *JAMA Oncol*, 3(4):524–548.

[53] Franceschini, G. et al. (2008). Conservative and radical oncoplastic approaches in the surgical treatment of breast cancer. *Eur Rev Med Pharmacol Sci*, 12(6):387–96.

[54] Friedman, J. (1989). Regularized discriminant analysis. *J Am Stat Assoc*, 84(405):165–175.

[55] Friedman, J. et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[56] Friedman, J. et al. (2009). GLMNET: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

[57] Gardner, E. (1988). The space of interactions in neural network models. *J Phys A-Math Gen*, 21(1):257.

[58] Geisser, S. (1964). Posterior odds for multivariate normal classifications. *J Roy Stat Soc B Met*, 26(1):69–76.

[59] Geisser, S. (1975). The predictive sample reuse method with applications. *J Am Stat Assoc*, 70(350):320–328.

[60] George, E. and McCulloch, R. (1997). Approaches for Bayesian variable selection. *Stat Sinica*, 7(2):339–373.

[61] Giles, D. et al. (2009). Bias of the maximum likelihood estimators of the two-parameter gamma distribution revisited. Technical report, Department of Economics, University of Victoria.

[62] Gradshteyn, I. and Ryzhik, I. (2007). *Table of Integrals, Series, and Products, D. Zwillinger, Ed.* Academic Press, Elsevier Inc.

[63] Greenland, S. et al. (2016). Sparse data bias: A problem hiding in plain sight. *BMJ*, 352:i1981.

[64] Grootendorst, M. et al. (2017). Use of a handheld terahertz pulsed imaging device to differentiate benign and malignant breast tissue. *Biomed Opt Express*, 8(6):2932–2945.

[65] Haff, L. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann Stat*, 8(3):586–597.

[66] Haka, A. et al. (2005). Diagnosing breast cancer by using Raman spectroscopy. *P Natl Acad Sci USA*, 102(35):12371–12376.

[67] Haldane, J. and Smith, S. (1956). The sampling distribution of a maximum-likelihood estimate. *Biometrika*, 43(1/2):96–103.

[68] Heider, F. (1946). Attitudes and cognitive organization. *J Psychol*, 21(1):107–112.

[69] Heimel, J. and Coolen, A. (2001). Supervised learning with restricted training sets: A generating functional analysis. *J Phys A-Math Gen*, 34(42):9009.

[70] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J Educ Psychol*, 24(6):417.

[71] Huang, J. and Harrington, D. (2002). Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics*, 58(4):781–791.

[72] Hubbard, J. (1959). Calculation of partition functions. *Phys Rev Lett*, 3(2):77.

[73] Ibrahim, J. et al. (2011). Bayesian models for gene expression with DNA microarray data. *J Am Stat Assoc*, 97(457):88–99.

[74] Jaynes, E. (1968). Prior probabilities. *IEEE T Syst Sci Cyb*, 4(3):227–241.

[75] Jaynes, E. (1986). Bayesian methods: General background.

[76] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *P Roy Soc Lond A Mat*, 186(1007):453–461.

[77] Jemal, A. et al. (2011). Global cancer statistics. *CA-Cancer J Clin*, 61(2):69–90.

[78] Jorns, J. et al. (2012). Intraoperative frozen section analysis of margins in breast conserving surgery significantly decreases reoperative rates: One-year experience at an ambulatory surgical center. *Am J Clin Pathol*, 138(5):657–669.

[79] Jung, S. et al. (2009). PCA consistency in high dimension, low sample size context. *Ann Stat*, 37(6B):4104–4130.

[80] Kalbfleisch, J. and Prentice, R. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

[81] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481.

[82] Keehn, D. (1965). A note on learning for Gaussian properties. *IEEE T Inform Theory*, 11(1):126–132.

[83] Keller, A. D. et al. (2000). Bayesian classification of DNA array expression data. Technical report, UW-CSE-2000-08-01, University of Washington.

[84] Kim, M. et al. (2016). The efficacy of intraoperative frozen section analysis during breast-conserving surgery for patients with ductal carcinoma in situ. *Breast cancer: Basic and clinical research*, 10:205.

[85] Klein, J. and Moeschberger, M. (2006). *Survival analysis: Techniques for censored and truncated data*. Springer Science & Business Media.

[86] Krzakala, F. et al. (2016). *Statistical physics, optimization, inference, and message-passing algorithms*. Oxford University Press.

[87] Li, B. and Saad, D. (2018). Exploring the function space of deep-learning machines. *Phys Rev Lett*, 120(24):248301.

[88] Liu, C.-H. et al. (2013). Resonance Raman and Raman spectroscopy for breast cancer detection. *Technol Cancer Res T*, 12(4):371–382.

[89] Livan, G. et al. (2018). *Introduction to random matrices: Theory and practice*. Springer.

[90] Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Stat Sinica*, 12(1):31–46.

[91] MacKay, D. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Comput*, 11(5):1035–1068.

[92] Mackinnon, M. and Puterman, M. (1989). Collinearity in generalized linear models. *Commun Stat Theory*, 18(9):3463–3472.

[93] Marquaridt, D. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612.

[94] Marčenko, V. and Pastur, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536.

[95] McCullagh, P. (2019). *Generalized linear models*. Routledge.

[96] Mézard, M. et al. (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company.

[97] Mézard, M. and Montanari, A. (2009). *Information, physics, and computation*. Oxford University Press.

[98] Mitrouli, M. and Roupa, P. (2018). Estimates for the generalized cross-validation function via an extrapolation and statistical approach. *Calcolo*, 55(3):24.

[99] Mosier, C. et al. (1951). Symposium: The need and means of cross-validation. *Educ Psychol Meas*, 11(1):5–11.

[100] Mozeika, A. et al. (2009). Computing with noise: Phase transitions in Boolean formulas. *Phys Rev Lett*, 103(24):248701.

[101] Nash, J., Varadhan, R., and Grothendieck, G. (2013). A replacement and extension of the optim () function. *R Package Version*, 8:7.

[102] Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *Comput J*, 7(4):308–313.

[103] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

[104] Newman, M. and Barkema, G. (1999). *Monte carlo methods in statistical physics*, volume 24. Oxford University Press.

[105] Obuchi, T. and Kabashima, Y. (2018). Accelerating cross-validation in multinomial logistic regression with L1-regularization. *J Mach Learn Res*, 19(1):2030–2059.

[106] Owen, A. (2007). Infinitely imbalanced logistic regression. *J Mach Learn Res*, 8(Apr):761–773.

[107] Peduzzi, P. et al. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48(12):1503–1510.

[108] Quenouille, M. (1949). Approximate tests of correlation in time-series. *J Roy Stat Soc B Met*, 11(1):68–84.

[109] Raudys, Š. and Young, D. (2004). Results in statistical discriminant analysis: A review of the former Soviet Union literature. *J Multivariate Anal*, 89(1):1–35.

[110] Rawlings, J. et al. (2001). *Applied regression analysis: A research tool*. Springer Science & Business Media.

[111] Raychaudhuri, S. et al. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, page 455. NIH Public Access.

[112] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[113] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev*, 65(6):386.

[114] Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY.

[115] Saad, D. (1994). Explicit symmetries and the capacity of multilayer neural networks. *J Phys A-Math Gen*, 27(8):2719.

[116] Sakorafas, G. (2001). Breast cancer surgery-historical evolution, current status and future perspectives. *Acta Oncol*, 40(1):5–18.

[117] Salehi, F. et al. (2019). The impact of regularization on high-dimensional logistic regression. In *Adv Neur In*, pages 12005–12015.

[118] Schnabel, F. et al. (2014). A randomized prospective study of lumpectomy margin assessment with use of marginprobe in patients with nonpalpable breast malignancies. *Ann Surg Oncol*, 21(5):1589–1595.

[119] Schwarze, H. (1993). Learning a rule in a multilayer neural network. *J Phys A-Math Gen*, 26(21):5781.

[120] Sei, T. (2014). Infinitely imbalanced binomial regression and deformed exponential families. *J Stat Plan Infer*, 149:116–124.

[121] Seung, H. et al. (1992). Statistical mechanics of learning from examples. *Phys Rev A*, 45(8):6056.

[122] Shalabi, A. et al. (2018). Bayesian clinical classification from high-dimensional data: Signatures versus variability. *Stat Methods Med Res*, 27(2):336–351.

[123] Sheikh, M. and Coolen, A. (2019). Analysis of overfitting in the regularized Cox model. *J Phys A-Math Theor*, 52(38):384002.

[124] Sheikh, M. and Coolen, A. (2020). Accurate Bayesian data classification without hyperparameter cross-validation. *J Classif*, 37(2):277–297.

[125] Shenton, L. and Bowman, K. (1963). Higher moments of a maximum-likelihood estimate. *J Roy Stat Soc B Met*, 25(2):305–317.

[126] Shenton, L. and Bowman, K. (1969). Maximum likelihood estimator moments for the 2-parameter gamma distribution. *Sankhyā Ser B*, 31(3/4):379–396.

[127] Srivastava, S. and Gupta, M. (2006). Distribution-based Bayesian minimum expected risk for discriminant analysis. In *Information Theory, 2006 IEEE International Symposium on*, pages 2294–2298. IEEE.

[128] Srivastava, S., Gupta, M., and Frigyik, B. (2007). Bayesian quadratic discriminant analysis. *J Mach Learn Res*, 8(6):1277–1305.

[129] St John, E. et al. (2017a). Diagnostic accuracy of intraoperative techniques for margin assessment in breast cancer surgery: A meta-analysis. *Ann Surg*, 265(2):300–310.

[130] St John, E. et al. (2017b). Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: Towards an intelligent knife for breast cancer surgery. *Breast Cancer Res*, 19(1):59.

[131] Stewart, G. (1987). Collinearity and least squares regression. *Stat Sci*, 2(1):68–84.

[132] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J Roy Stat Soc B Met*, 39(1):44–47.

[133] Sur, P. and Candès, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *P Natl Acad Sci USA*, 116(29):14516–14525.

[134] Sur, P. et al. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab Theory Rel*, 175(1-2):487–558.

[135] Therneau, T. and Lumley, T. (2017). Package 'survival'. *R package version*, pages 2–41.

[136] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J Roy Stat Soc B Met*, 58(1):267–288.

[137] Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29(2):614.

[138] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press.

[139] Vittinghoff, E. and McCulloch, C. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*, 165(6):710–718.

[140] Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

[141] Wallace, B. et al. (2011). Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, pages 754–763. IEEE.

[142] West, A. (1997). Role of biases in neural network models. *Doctoral Dissertation*.

[143] West, M. et al. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *P Natl Acad Sci USA*, 98(20):11462–11467.

[144] Wilson, J. and Lorenz, K. (2015). Short history of the logistic regression model. In *Modeling Binary Correlated Responses using SAS, SPSS and R*, pages 17–23. Springer.

[145] Witten, D. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Stat Methods Med Res*, 19(1):29–51.

[146] Yang, R. and Berger, J. (1996). *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University.

[147] Yun, Y.-H. et al. (2014). A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration. *Anal Chim Acta*, 807:36–43.

[148] Zhang, H. et al. (2012). Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC Bioinformatics*, 13(1):1.

[149] Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.

Appendix A

Mathematical identities

A.1 Gaussian distribution results

A.1.1 Gaussian normalization

Consider a multivariate Gaussian distribution with mean, μ, and precision matrix, $\mathbf{A}$. The normalization condition is

$$\int_{\mu \in \mathbb{R}^p} \mathrm{d}\mu \; \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \, e^{-\frac{1}{2}\mu \mathbf{A} \mu} = 1 \tag{A.1}$$

A.1.2 Moments of the multivariate Gaussian distribution

The moment generating function of a random variable X is defined with $t \in \mathrm{R}$ as

$$M_X(t) \equiv \mathbb{E}(e^{tX}) \tag{A.2}$$

Similarly for a p-dimensional random variable X with $\mathbf{t} \in \mathrm{R}^p$

$$M_X(\mathbf{t}) \equiv \mathbb{E}(e^{\mathbf{t}.X}) \tag{A.3}$$

For a univariate random variable, $X \sim \mathrm{N}(\mu, \sigma^2)$, the moment generating functions becomes

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{\mathrm{dx}}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2 + tx} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{\mathrm{dx}}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}[x-(\mu+\sigma^2 t)]^2} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned} \tag{A.4}$$

A.2 Wishart distribution results

A.2.1 Wishart normalization

Assume a p-dimensional vector $\mathbf{x_i} \sim \mathrm{N}_p(0, \mathbf{S})$. The $p \times p$ symmetric matrix, $\Lambda = \frac{1}{r}\sum_{i=1}^{r} \mathbf{x_i}\mathbf{x_i}^T$ has a Wishart distribution with r degrees of freedom and scale matrix, $\mathbf{S}$.

$$\int_{\Lambda \succ 0} \frac{\mathrm{d}\Lambda\ |\Lambda|^{(\mathrm{r-p-1})/2}}{c(p,r)|\mathbf{S}|^{r/2}}\ e^{-\frac{1}{2}\mathrm{Tr}\Lambda\mathbf{S}^{-1}} = 1 \tag{A.5}$$

where the integral is over all positive definite matrices, $\Lambda_z \succ 0$ and the normalization term $c(p,r) = 2^{rp/2}\ \Gamma_p(r/2)$. The degrees of freedom, $r > p-1$ are not restricted to integer values i.e. $r \in \mathbb{R}$.

A.2.2 Conjugate prior

The Wishart distribution is the conjugate prior to the multivariate Gaussian distribution with unknown precision matrix (inverse of covariance matrix). If instead, we make the assumption $\mathbf{x_i} \sim \mathrm{N}_p(\mu, \mathbf{S})$, our covariance matrix Λ will have a non-Central Wishart distribution with an additional non-centrality parameter to be estimated. Since the prior distribution does not assume any knowledge of the data, the Central Wishart distribution, as defined in (A.5), is sufficient for our inference problem.

A.2.3 Moments of the multivariate Gaussian distribution

To find the moments of the Wishart distribution, we use the generalization of the characteristic function to real-valued matrices with argument $\Theta \in \mathbb{R}^{p\times p}$

$$\phi_\Lambda(\Theta) = \mathbb{E}\left[e^{\mathrm{i}\ \mathrm{Tr}(\Theta\Lambda)}\right] \tag{A.6}$$

where the expectation is over the probability density function of Λ. This expression is the Fourier transform of the probability density function and an explicit form is determined using the normalization (A.5).

$$\phi_\Lambda(\Theta) = |\mathbb{I}_p - 2\mathrm{i}\ \Theta\ \mathbf{S}|^{-r/2} \tag{A.7}$$

Taking the first derivative with respect to the matrix Θ using the identity

$$\frac{\partial}{\partial \mathbf{M}}|\mathbf{M}| = |\mathbf{M}|\ (\mathbf{M}^{-1})^T \tag{A.8}$$

results in

$$\frac{\partial}{\partial \Theta}\phi(\Theta) = \mathrm{i}r\mathbf{S}\,|\mathbb{I}_p - 2\mathrm{i}\,\Theta\,\mathbf{S}|^{-r/2}\,(\mathbb{I}_p - 2\mathrm{i}\,\Theta\,\mathbf{S})^{-1} \tag{A.9}$$

The moments are calculated via

$$\mathbb{E}(\Lambda^k) = \mathrm{i}^{-k}\left[\frac{\partial^k \phi(\Theta)}{\partial \Theta}\right]_{\Theta=\mathbf{0}} \tag{A.10}$$

For $k = 1$,

$$\mathbb{E}(\Lambda) = r\mathbf{S} \tag{A.11}$$

A.3 Hubbard Stratonovich transformation

This transformation [72] is used to linearize expressions with interacting variables by introducing a field variable x.

$$\sqrt{\frac{1}{2\pi a}}\int_{-\infty}^{\infty} \mathrm{d}y\; e^{-\frac{1}{2a}y^2 - ixy} = e^{-\frac{1}{2}ax^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi a}}\; e^{-\frac{1}{2a}(y+\mathrm{i}ax)^2} = e^{-\frac{1}{2}ax^2} \tag{A.12}$$

A.4 Integral representation of Dirac delta function

Using the following property $f(x_0) = \int \mathrm{d}x\; \delta(x - x_0)$ and taking the Fourier transform of the Dirac delta function

$$\hat{\delta}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}t\;\delta(t-t_0)\;e^{i\omega t} = \frac{1}{\sqrt{2\pi}}e^{i\omega t_0} \tag{A.13}$$

Now take the inverse Fourier transform to achieve the required result

$$\delta(t-t_0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}\omega\;\hat{\delta}(w)\;e^{-i\omega t} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}\omega\;\frac{1}{\sqrt{2\pi}}e^{i\omega t_0}e^{-i\omega t} \tag{A.14}$$

Hence we find the integral representation of the Dirac delta function.

$$\delta(t-t_0) = \int_{-\infty}^{\infty}\frac{\mathrm{d}\omega}{2\pi}\;e^{i\omega(t_0-t)} \tag{A.15}$$

where $t, t_0 \in \mathbb{R}$.

For comparison the Kronecker delta function which admits integer inputs can be represented by

$$\delta_{n,m} = \int_{-\pi}^{\pi} \mathrm{d}\omega \; e^{i\omega(n-m)} \tag{A.16}$$

Note the different limits of integration.

A.5 The replica identity

$$\begin{aligned}
\lim_{n\to 0} \frac{1}{n} \log\langle Z^n \rangle &= \lim_{n\to 0} \frac{1}{n} \log\langle e^{n \log Z} \rangle \\
&= \lim_{n\to 0} \frac{1}{n} \log\left\langle 1 + n \log Z + \mathscr{O}(n^2) \right\rangle \qquad \text{Taylor expansion of exponential} \\
&= \lim_{n\to 0} \frac{1}{n} \log\left[1 + n\langle \log Z \rangle + \mathscr{O}(n^2) \right] \qquad \text{Linearity of expectation} \\
&= \langle \log Z \rangle \qquad\qquad\qquad \text{Re-exponentiation}
\end{aligned} \tag{A.17}$$

A.6 Lambert W-function

$W(x)$ denotes Lambert's W-function. It is the inverse of $f(x) = x\mathrm{e}^x$ and is used extensively in the analysis of the regularized Cox model. The following identities are useful:

$$\frac{\mathrm{dW(z)}}{\mathrm{dz}} = \frac{W(z)}{z[1+W(z)]} \tag{A.18}$$

$$W(z) = z + O(z^2) \quad \text{for } \mathrm{z} \to 0 \tag{A.19}$$

$$W(z)\, e^{W(z)} = z \tag{A.20}$$

A.7 Confusion matrix

Confusion matrices summarize classification results. Assume a binary classification problem. Common terms in medical statistics can be defined in terms of TP, TN, FP and FN.

	Predict positive	Predict negative
True class positive	True positive (TP)	False negative (FN)
True class negative	False positive (FP)	True negative (TN)

Table A.1 Summary of confusion matrix format displaying true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{A.21a}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{A.21b}$$

$$\text{Positive Predictive Value } (\text{PPV}) = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{A.21c}$$

$$\text{Negative Predictive Value } (\text{NPV}) = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{A.21d}$$

Appendix B

Supplementary Replica Calculations

B.1 Transformation in the unregularized case

We show that the covariance matrix $\mathbf{A}$ can be transformed away in the maximum likelihood case $\eta = 0$ for the logistic regression model. We transform $\mathbf{z}_i = \langle \mathbf{z} \rangle + \mathbf{A}^{\frac{1}{2}} \tilde{\mathbf{z}}_i$ where $A_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} (z_{i\mu} - \langle z_\mu \rangle)(z_{i\nu} - \langle z_\nu \rangle)$. By maximizing the log likelihood for the three models, we find the inferred regression coefficients do not depend on the form of $\mathbf{A}$.

$$\begin{aligned}
\hat{\beta} &= \text{argmax}_\beta \sum_{i=1}^{N} [t_i(r + \beta \cdot \mathbf{z}_i) - \log 2\cosh(r + \beta \cdot \mathbf{z}_i)] \\
&= \text{argmax}_\beta \sum_{i=1}^{N} \left[t_i(r + \beta \cdot \langle \mathbf{z} \rangle + \beta \cdot \mathbf{A}^{\frac{1}{2}} \tilde{\mathbf{z}}_i) - \log 2\cosh(r + \beta \cdot \langle \mathbf{z} \rangle + \beta \cdot \mathbf{A}^{\frac{1}{2}} \tilde{\mathbf{z}}_i) \right] \\
&= \text{argmax}_\beta \sum_{i=1}^{N} \left[t_i(r + (\mathbf{A}^{\frac{1}{2}} \beta) \cdot \tilde{\mathbf{z}}_i) - \log 2\cosh(r + (\mathbf{A}^{\frac{1}{2}} \beta) \cdot \tilde{\mathbf{z}}_i) \right]
\end{aligned} \tag{B.1}$$

So $\hat{\beta} = \mathbf{A}^{-\frac{1}{2}} \tilde{\beta}$ where $\tilde{\beta}$ is the inferred regression vector with respect to the zero mean, uncorrelated covariates $\tilde{\mathbf{z}}_i$.

$$\tilde{\beta} = \text{argmax}_\beta \, [t_i(r + \beta \cdot \tilde{\mathbf{z}}_i) - \log 2\cosh(r + \beta \cdot \tilde{\mathbf{z}}_i)] \tag{B.2}$$

Therefore, in the absence of regularization, we can transform correlated covariates into uncorrelated ones and proceed with maximum likelihood in the usual manner. The same logic applied for linear and Cox regression [26].

B.2 Integral over regression coefficients

To transform the β integral in (2.33) into a Gaussian integral, we define $\Xi \in \mathbb{R}^{np\times np}$ and $\xi \in \mathbb{R}^{np}$.

$$\Xi_{\alpha\mu;\beta\nu} \equiv 2\eta\gamma\delta_{\alpha\beta}(\mathbf{A}^{-1})_{\mu\nu} + 2i\delta_{\mu\nu}\hat{C}_{\alpha\beta} \text{ and } \xi_\mu^\alpha \equiv -2i\hat{C}_{0\alpha}\tilde{\beta}_\mu^0 \tag{B.3}$$

where the $\tilde{\beta}$ vector is composed of the n replicas of the regression coefficient vectors arranged in $1\times np$ vector

$$\tilde{\beta} \equiv \{\tilde{\beta}_1^1,\ldots,\tilde{\beta}_p^1,\tilde{\beta}_1^2,\ldots,\tilde{\beta}_p^2,\ldots,\tilde{\beta}_1^n,\ldots,\tilde{\beta}_p^n\} \tag{B.4}$$

Similarly the $1\times np$ vector ξ

$$\xi \equiv \{\xi_1^1,\ldots,\xi_p^1,\xi_1^2,\ldots,\xi_p^2,\ldots,\xi_1^n,\ldots,\xi_p^n\} \tag{B.5}$$

The β integral can be transformed into a np dimensional Gaussian integral and therefore evaluated.

$$\begin{aligned}
&\int\left(\prod_{\alpha=1}^{n} d\tilde{\beta}^\alpha e^{-\eta\gamma\tilde{\beta}^\alpha\cdot\mathbf{A}^{-1}\tilde{\beta}^\alpha}\right)e^{-i\sum_{\alpha,\rho=1}^{n}\hat{C}_{\alpha\rho}\tilde{\beta}^\alpha\cdot\tilde{\beta}^\rho - 2i\sum_{\rho=1}^{n}\hat{C}_{0\rho}\tilde{\beta}^0\cdot\tilde{\beta}^\rho} \\
&= \int \mathrm{d}\tilde{\beta}\; e^{-\frac{1}{2}\sum\limits_{\alpha,\rho=1}^{n}\sum\limits_{\mu,\nu=1}^{p}\tilde{\beta}_\mu^\alpha\tilde{\beta}_\nu^\rho\left[2\eta\gamma\delta_{\alpha\rho}(\mathbf{A}^{-1})_{\mu\nu}+2i\delta_{\mu\nu}\hat{C}_{\alpha\rho}\right]-2i\sum\limits_{\rho=1}^{n}\sum\limits_{\mu=1}^{p}\tilde{\beta}_\mu^\rho\hat{C}_{0\rho}\tilde{\beta}_\mu^0} \\
&= \int d\tilde{\beta}\, e^{-\frac{1}{2}\tilde{\beta}^T\cdot\Xi\tilde{\beta}+\xi\cdot\tilde{\beta}} = e^{\frac{1}{2}\xi^T\cdot\Xi^{-1}\xi}\int d\tilde{\beta}\, e^{-\frac{1}{2}(\tilde{\beta}-\Xi^{-1}\xi)^T\cdot\Xi(\tilde{\beta}-\Xi^{-1}\xi)} = \frac{(2\pi)^{\frac{np}{2}}}{|\Xi|^{\frac{1}{2}}}e^{\frac{1}{2}\xi^T\cdot\Xi^{-1}\xi}
\end{aligned} \tag{B.6}$$

It will be useful to show the two parts of $\Xi_{\alpha\mu;\beta\nu} \equiv 2\eta\gamma\delta_{\alpha\beta}(\mathbf{A}^{-1})_{\mu\nu} + \delta_{\mu\nu}D_{\alpha\beta}$ commute.

$$\begin{aligned}
\sum_{r=1}^{n}\sum_{s=1}^{p}(P)_{ij;rs}(Q)_{rs;kl} &= \sum_{r=1}^{n}\sum_{s=1}^{p}2\eta\gamma\delta_{ir}(\mathbf{A}^{-1})_{js}\delta_{sl}D_{rk} = 2\eta\gamma D_{ik}(\mathbf{A}^{-1})_{jl} \\
\sum_{r=1}^{n}\sum_{s=1}^{p}(Q)_{ij;rs}(P)_{rs;kl} &= \sum_{r=1}^{n}\sum_{s=1}^{p}\delta_{js}D_{ir}2\eta\gamma\delta_{rk}(\mathbf{A}^{-1})_{sl} = 2\eta\gamma D_{ik}(\mathbf{A}^{-1})_{jl}
\end{aligned} \tag{B.7}$$

B.3 Replica symmetric simplifications

We simplify two terms resulting from β integral in (2.37) using the replica symmetric ansatz. One eigenvector of $\mathbf{D}$ is $\mathbf{u} = (1,1,\ldots,1)^T$ with the corresponding eigenvalue $D+(n-1)d$.

The other $n-1$ eigenvectors satisfy $\sum_{i=1}^{n}\mathbf{u}_i = 0$ with corresponding eigenvalue $D-d$. For $\mathbf{D}$ to be positive definite, $D > d$. Let $\{a_\mu\}$ and $\{b_\alpha\}$ denote the eigenvalues of $\mathbf{A}$ and $\mathbf{D}$ and evaluate

$$\begin{aligned}\frac{1}{2N}\log\det\Xi &= \frac{1}{2N}\sum_{\mu=1}^{p}\sum_{\alpha=1}^{n}\log\left(\frac{2\eta\gamma}{a_\mu}+b_\alpha\right)\\ &= \frac{1}{2}\zeta\frac{1}{p}\sum_{\mu=1}^{p}\left\{\log\left(\frac{2\eta\gamma}{a_\mu}+D+(n-1)d\right)+(n-1)\log\left(\frac{2\eta\gamma}{a_\mu}+(D-d)\right)\right\}\end{aligned} \tag{B.8}$$

Using (2.34), the transformation $\tilde{\beta} = \mathbf{A}^{\frac{1}{2}}\beta$ and considering the large p limit, we can evaluate the term

$$\begin{aligned}-\frac{1}{2N}\xi^T\Xi^{-1}\xi &= -\frac{1}{2}\zeta\frac{1}{p}\sum_{\alpha,\beta=1}^{n}\sum_{\mu,\nu=1}^{p}\xi_\mu^\alpha\Xi^{-1}_{\alpha\mu;\beta\nu}\xi_\nu^\beta\\ &= -\frac{1}{2}\zeta\frac{1}{p}\sum_{\alpha,\beta=1}^{n}\sum_{\mu,\nu=1}^{p}(-d_\alpha\tilde{\beta}^0_\mu)\Xi^{-1}_{\alpha\mu;\beta\nu}(-d_\beta\tilde{\beta}^0_\nu)\\ &= -\frac{1}{2}\zeta d_0^2\frac{1}{p}\sum_{\mu,\nu}\left(\sum_{\mu'=1}^{p}(\mathbf{A}^{\frac{1}{2}})_{\mu\mu'}\beta^0_{\mu'}\right)\left(\sum_{\nu'=1}^{p}(\mathbf{A}^{\frac{1}{2}})_{\nu\nu'}\beta^0_{\nu'}\right)\sum_{\alpha,\beta=1}^{n}\Xi^{-1}_{\alpha\mu;\beta\nu}\\ &= -\frac{1}{2}\zeta d_0^2\frac{1}{p}\sum_{\alpha,\beta=1}^{n}\sum_{\mu',\nu'=1}^{p}\beta^0_{\mu'}\beta^0_{\nu'}\underbrace{\sum_{\mu,\nu}(\mathbf{A}^{\frac{1}{2}})_{\mu\mu'}(\mathbf{A}^{\frac{1}{2}})_{\nu\nu'}\Xi^{-1}_{\alpha\mu;\beta\nu}}_{P_{\mu'\nu'}}\\ &= -\frac{1}{2}\zeta d_0^2\sum_{\alpha,\beta=1}^{n}\frac{S^2}{p}\sum_{\mu'=1}^{p}P_{\mu'\mu'} = -\frac{1}{2}\zeta d_0^2S^2n\int\frac{\mathrm{d}a\,\rho(a)\,a^2}{2\eta\gamma+[D+(n-1)d]a}\end{aligned} \tag{B.9}$$

where $\mathbf{P}$ is a $p\times p$ matrix and the penultimate equality uses the self-averaging argument in the following Appendix B.4.

B.4 Self-averaging with respect to true associations

Here we investigate properties of random variables of the form $\mathscr{R} = p^{-1}\beta^0\cdot\mathbf{P}\beta^0$ in the limit $p\to\infty$, where the true association vectors $\beta^0 = \{\beta^0_\mu\}$ are drawn randomly from some distribution $p(\beta^0)$ and $\mathbf{P}$ is a fixed symmetric positive definite $p\times p$ matrix, which is independent of β^0. In particular, we wish to determine under which conditions $\mathscr{R}$ will be self-averaging, i.e. $\lim_{p\to\infty}\langle\mathscr{R}\rangle > 0$ exists, and $\lim_{p\to\infty}[\langle\mathscr{R}^2\rangle - \langle\mathscr{R}\rangle^2] = 0$. Brackets will in

this Appendix denote averaging over $p(\beta^0)$, and we will write the eigenvalue distribution of $\mathbf{P}$ as $\rho(\lambda)$. We make the following assumptions:[1]

1. The $\{\beta_\mu^0\}$ are independent and identically distributed, i.e. $p(\beta^0) = \prod_{\mu=1}^p p(\beta_\mu^0)$.
2. $p(\beta_\mu^0)$ is symmetric in β_μ^0, with finite second and fourth order moments.

In view of our earlier definition $S^2 = \lim_{p\to\infty} p^{-1}(\beta^0)^2$, we must identify $\langle(\beta_\mu^0)^2\rangle = S^2$. We will write $\Sigma = \langle(\beta_\mu^0)^4\rangle$. It then follows that

$$\langle\mathscr{R}\rangle_{\mathscr{D}} \equiv \lim_{p\to\infty}\left\langle \frac{1}{p}\sum_{\mu,\nu=1}^{p}\beta_\mu^0\beta_\nu^0 P_{\mu\nu}\right\rangle_{\mathscr{D}} = \frac{1}{p}\sum_{\mu,\nu=1}^{p}\langle\beta_\mu^0\beta_\nu^0\rangle_{\mathscr{D}}P_{\mu\nu} = \frac{S^2}{p}\sum_{\mu=1}^{p}P_{\mu\mu} \sim \mathscr{O}(1) \tag{B.10}$$

The second term is slightly more involved and we explicitly write out the various combinations of indices

$$\begin{aligned}\langle\mathscr{R}^2\rangle_{\mathscr{D}} &= \lim_{p\to\infty}\left\langle\left(\frac{1}{p}\sum_{\mu,\nu=1}^{p}\beta_\mu^0\beta_\nu^0P_{\mu\nu}\right)\left(\frac{1}{p}\sum_{\kappa,\tau=1}^{p}\beta_\kappa^0\beta_\tau^0P_{\kappa\tau}\right)\right\rangle_{\mathscr{D}}\\ &= \lim_{p\to\infty}\frac{1}{p^2}\sum_{\mu,\nu,\kappa,\tau}\langle\beta_\mu^0\beta_\nu^0\beta_\kappa^0\beta_\tau^0\rangle_{\mathscr{D}}P_{\mu\nu}P_{\kappa\tau}\end{aligned} \tag{B.11}$$

We now explicitly consider each of the four cases where $\langle\beta_\mu^0\beta_\nu^0\beta_\kappa^0\beta_\tau^0\rangle_{\mathscr{D}}$ has non-zero values. For clarity, the notation $P_{\mu\mu}^2$ means the square of entry $P_{\mu\mu}$ an entry of the matrix P squared.

1. $\mu=\nu,\ \kappa=\tau,\ \mu\neq\kappa \quad \Rightarrow \delta_{\mu\nu}\delta_{\kappa\tau}(1-\delta_{\mu\kappa})$

$$\begin{aligned}\lim_{p\to\infty}\frac{1}{p^2}\sum_{\mu,\nu,\kappa,\tau}\langle\beta_\mu^0\,\beta_\nu^0\beta_\kappa^0\,\beta_\tau^0\rangle P_{\mu\nu}P_{\kappa\tau} &= \lim_{p\to\infty}\frac{S^4}{p^2}\sum_{\mu,\nu,\kappa,\tau}\delta_{\mu\nu}\delta_{\kappa\tau}(1-\delta_{\mu\kappa})P_{\mu\nu}P_{\kappa\tau}\\ &= \lim_{p\to\infty}\frac{S^4}{p^2}\left\{\sum_{\mu,\nu,\kappa,\tau}\delta_{\mu\nu}\delta_{\kappa\tau}P_{\mu\nu}P_{\kappa\tau} - \delta_{\mu\nu}\delta_{\kappa\tau}\delta_{\mu\kappa}P_{\mu\nu}P_{\kappa\tau}\right\}\\ &= \lim_{p\to\infty}\frac{S^4}{p^2}\left\{\sum_{\mu,\kappa}P_{\mu\mu}P_{\kappa\kappa} - \sum_{\mu}P_{\mu\mu}P_{\mu\mu}\right\}\\ &= \lim_{p\to\infty}\frac{S^4}{p^2}\left\{\left(\sum_{\mu}P_{\mu\mu}\right)\left(\sum_{\kappa}P_{\kappa\kappa}\right) - \sum_{\mu}P_{\mu\mu}^2\right\}\\ &= \lim_{p\to\infty}\frac{S^4}{p^2}\left\{\left(\sum_{\mu}P_{\mu\mu}\right)^2 - \sum_{\mu}P_{\mu\mu}^2\right\}\end{aligned} \tag{B.12}$$

[1] Assuming distinct variances for each β_μ^0 complicates various equations but ultimately leads to similar final conditions on the eigenvalue spectrum of $\mathbf{A}$.

2. $\mu=\kappa,\ \nu=\tau,\ \mu\neq\nu \quad \Rightarrow \delta_{\mu\kappa}\delta_{\nu\tau}(1-\delta_{\mu\nu})$

$$\lim_{p\to\infty}\frac{1}{p^2}\sum_{\mu,\nu,\kappa,\tau}\langle\beta^0_\mu\beta^0_\nu\beta^0_\kappa\beta^0_\tau\rangle P_{\mu\nu}P_{\kappa\tau}=\lim_{p\to\infty}\frac{S^4}{p^2}\left\{\sum_{\mu,\nu}P^2_{\mu\nu}-\sum_{\mu}P^2_{\mu\mu}\right\} \tag{B.13}$$

3. $\mu=\tau,\ \nu=\kappa,\ \mu\neq\nu \quad \Rightarrow \delta_{\mu\tau}\delta_{\nu\kappa}(1-\delta_{\mu\nu})$

$$\lim_{p\to\infty}\frac{1}{p^2}\sum_{\mu,\nu,\kappa,\tau}\langle\beta^0_\mu\beta^0_\nu\beta^0_\kappa\beta^0_\tau\rangle P_{\mu\nu}P_{\kappa\tau}=\lim_{p\to\infty}\frac{S^4}{p^2}\left\{\sum_{\mu,\nu}P^2_{\mu\nu}-\sum_{\mu}P^2_{\mu\mu}\right\} \tag{B.14}$$

4. $\mu=\nu=\kappa=\tau \quad \Rightarrow \delta_{\mu\nu}\delta_{\nu\kappa}\delta_{\kappa\tau}$

$$\lim_{p\to\infty}\frac{1}{p^2}\sum_{\mu,\nu,\kappa,\tau}\langle\beta^0_\mu\beta^0_\nu\beta^0_\kappa\beta^0_\tau\rangle P_{\mu\nu}P_{\kappa\tau}=\lim_{p\to\infty}\frac{\langle(\beta^0)^4\rangle}{p^2}\left\{\sum_{\mu}P^2_{\mu\mu}\right\} \tag{B.15}$$

Putting these results together

$$\begin{aligned}\langle\mathscr{R}^2\rangle_{\mathscr{D}}&=\lim_{p\to\infty}\frac{S^4}{p^2}\left\{\left(\sum_\mu P_{\mu\mu}\right)^2+2\sum_{\mu,\nu}P^2_{\mu\nu}+\left(\frac{\langle(\beta^0)^4\rangle_{\mathscr{D}}}{S^4}-3\right)\sum_\mu P^2_{\mu\mu}\right\}\\&=\lim_{p\to\infty}\frac{S^4}{p^2}\left\{p^2\left(\frac{1}{p}\sum_\mu P_{\mu\mu}\right)^2+2\sum_{\mu,\nu}P^2_{\mu\nu}+\left(\frac{\langle(\beta^0)^4\rangle_{\mathscr{D}}}{S^4}-3\right)\sum_\mu P^2_{\mu\mu}\right\}\\&=\langle\mathscr{R}\rangle^2_{\mathscr{D}}+\lim_{p\to\infty}\left\{\frac{2S^4}{p^2}\sum_{\mu,\nu}P^2_{\mu\nu}+\left(\langle(\beta^0)^4\rangle_{\mathscr{D}}-3S^4\right)\frac{1}{p^2}\sum_\mu P^2_{\mu\mu}\right\}\end{aligned} \tag{B.16}$$

We conclude that $\mathscr{R}$ will be self-averaging in the limit $p\to\infty$ if $\lim_{p\to\infty}p^{-1}\sum_{\mu=1}^p P_{\mu\mu}\in\mathbb{R}$ and $\lim_{p\to\infty}p^{-2}\sum_{\mu,\nu=1}^p P^2_{\mu\nu}=0$ or equivalently[2]

$$\lim_{p\to\infty}\int \mathrm{d}\lambda\ \rho(\lambda)\lambda\ \text{ exists and }\quad \lim_{\mathrm{p}\to\infty}\mathrm{p}^{-1}\int\mathrm{d}\lambda\ \rho(\lambda)\lambda^2=0 \tag{B.17}$$

where $\{\lambda_\mu\}_{\mu=1}^p$ are the eigenvalues of $\mathbf{P}$.

The two relevant quadratic expression for which we seek to demonstrate self-averaging are the following:

[2] $(\mathbf{P}^2)_{\mu\mu}=\sum_{\nu=1}^p P_{\mu\nu}P_{\nu\mu}=\sum_{\nu=1}^p P^2_{\mu\nu} \quad\Rightarrow\quad \mathrm{Tr}(\mathbf{P}^2)=\sum_{\mu=1}^p(\mathbf{P}^2)_{\mu\mu}=\sum_{\mu,\nu=1}^p P^2_{\mu\nu}$

1. Application to $\mathbf{P}=\mathbf{A}$ tells us that if $\lim_{p\to\infty}\langle a\rangle\in\mathbb{R}$ and $\lim_{p\to\infty}p^{-1}\langle a^2\rangle=0$ (i.e. the covariate correlations are not excessive), then

$$\tilde{S}^2 \;=\; \lim_{p\to\infty}\frac{1}{p}\beta^0\cdot\mathbf{A}\beta^0=S^2\langle a\rangle \tag{B.18}$$

2. Our second application is to the following matrix, in which the vectors $\{\mathbf{v}^\mu\}$ are the orthogonal and normalised eigenvectors of $\mathbf{A}$, with eigenvalues a_μ:

$$P_{\mu\nu} \;=\; \sum_{\rho=1}^{p}\frac{a_\rho^2 v_\mu^\rho v_\nu^\rho}{2\eta\gamma+ga_\rho} \tag{B.19}$$

Here we find, anticipating that $g>0$ and using $\eta\gamma>0$,

$$\frac{1}{p}\sum_{\mu=1}^{p}P_{\mu\mu} \;=\; \frac{1}{p}\sum_{\rho=1}^{p}\frac{a_\rho^2}{2\eta\gamma+ga_\rho}\leq\frac{\langle a\rangle}{g} \tag{B.20}$$

$$\begin{aligned}\frac{1}{p^2}\sum_{\mu\nu=1}^{p}P_{\mu\nu}^2 \;&=\; \frac{1}{p^2}\sum_{\rho\rho'\mu\nu=1}^{p}\frac{a_\rho^2a_{\rho'}^2v_\mu^\rho v_\nu^\rho v_\mu^{\rho'}v_\nu^{\rho'}}{(2\eta\gamma+ga_\rho)(2\eta\gamma+ga_{\rho'})}\\ &=\; \frac{1}{p^2}\sum_{\rho=1}^{p}\frac{a_\rho^4}{(2\eta\gamma+ga_\rho)^2}\leq\frac{\langle a^2\rangle}{pg^2}\end{aligned} \tag{B.21}$$

We conclude, provided $g>0$, that the same two conditions on $\mathbf{A}$ that guarantee self-averaging of $\tilde{S}^2$ for $p\to\infty$ will also imply self-averaging here:

$$\lim_{p\to\infty}\frac{1}{p}\sum_{\rho=1}^{p}\frac{a_\rho^2(\beta^0\cdot\mathbf{v}^\rho)^2}{2\eta\gamma+ga_\rho} \;=\; \Big\langle\frac{S^2a^2}{2\eta\gamma+ga}\Big\rangle \tag{B.22}$$

Thus, for our RS theory to be self-averaging with respect to the realisation of the true association vector β^0 (given our mild assumptions on the distribution from which β^0 is drawn), it is sufficient that average and width of the eigenvalue distribution $\rho(a)$ of the covariate correlation matrix $\mathbf{A}$ remain finite in the limit $p\to\infty$.

B.5 Convexity of overfitting measure

A function $f(x)$ is convex over a region $\mathbf{X}$ if $\forall x_1,x_2\in\mathbf{X}$ and $\forall\alpha\in[0,1]$

$$f\left[\alpha x_1+(1-\alpha)x_2\right] \;<\; \alpha f(x_1)+(1-\alpha)f(x_2) \tag{B.23}$$

Applying this inequality to the overfitting measure (3.3) results in the following convexity condition

$$\frac{1}{N}\sum_{i=1}^{N}\log\cosh\left[(\alpha\beta_1+(1-\alpha)\beta_2)\cdot\mathbf{z}_i\right] < \frac{1}{N}\sum_{i=1}^{N}\left[\alpha\log\cosh\beta_1\cdot\mathbf{z}_i+(1-\alpha)\log\cosh\beta_2\cdot\mathbf{z}_i\right] \tag{B.24}$$

Since cosh is a convex function and summations preserve convexity, we find E is a convex function.

B.6 Calculation of Marčenko-Pastur integrals

B.6.1 Mean eigenvalue

We consider the normalized covariance matrix $N^{-1}\mathbf{Z}\mathbf{Z}^T$ and repeatedly use indefinite integrals from sections 2.261-2 of [62] . The mean value of the empirical eigenvalue distribution is

$$I=\int_{\alpha}^{\beta} x\,\frac{\mathrm{d}x}{2\pi\zeta}\,\frac{\sqrt{(x-\alpha)(\beta-x)}}{x}=\int_{\alpha}^{\beta}\frac{\mathrm{d}x}{2\pi\zeta}\,\sqrt{R} \tag{B.25}$$

where $R=(x-\alpha)(\beta-x)$ and the limits of integration $\alpha=\left(1-\sqrt{\zeta}\right)^2$ and $\beta=\left(1+\sqrt{\zeta}\right)^2$. These two definite integrals will be useful

$$\int_{\alpha}^{\beta}\frac{\mathrm{dx}}{\sqrt{R}}=\pi \quad\text{and}\quad \int_{\alpha}^{\beta}\frac{\mathrm{dx}}{\mathrm{x}\sqrt{\mathrm{R}}}=\frac{\pi}{\sqrt{\alpha\beta}} \tag{B.26}$$

Using equation 2.262 [62]

$$\int \mathrm{d}x\,\sqrt{R}=\frac{[2x-(\alpha+\beta)]\sqrt{R}}{4}+\frac{(\alpha-\beta)^2}{8}\int\frac{\mathrm{d}x}{\sqrt{R}} \tag{B.27}$$

Inserting expressions for the limits α and β produces the normalization condition, $\mathrm{I}=1$.

B.6.2 Required Integral

We are left to calculate an integral of this form

$$\mathrm{I}=\int_{\alpha}^{\beta}\mathrm{dx}\,\frac{\sqrt{(\mathrm{x}-\alpha)(\beta-\mathrm{x})}}{2\pi\mathrm{x}^2} \tag{B.28}$$

with limits $\alpha = \left(1-\sqrt{\zeta}\right)^2$ and $\beta = \left(1+\sqrt{\zeta}\right)^2$. Using equation 2.267 [62]

$$\int \frac{\mathrm{dx}\ \sqrt{\mathrm{R}}}{x^2} = -\frac{\sqrt{R}}{x} + \frac{\alpha+\beta}{2}\int \frac{\mathrm{dx}}{x\sqrt{R}} - \int \frac{\mathrm{dx}}{\sqrt{R}} \tag{B.29}$$

Noting the first term disappears at the two limits and evaluating the definite integrals for the second and third terms The definite integral can be calculated using (B.26)

$$\int_{\alpha}^{\beta} \frac{\mathrm{dx}\ \sqrt{\mathrm{R}}}{2\pi x^2} = \frac{1}{2\pi}\left[\frac{\alpha+\beta}{2}\frac{\pi}{\sqrt{\alpha\beta}} - \pi\right] = \frac{(\alpha+\beta)-2\sqrt{\alpha\beta}}{4\sqrt{\alpha\beta}} \tag{B.30}$$

Note this result was also found in [89]. The required result is found using explicit expressions for α and β in terms of ζ

$$I = \frac{\zeta}{1-\zeta} \tag{B.31}$$

B.7 Cox proportional hazards relationships

The distribution of event times can be defined via the probability density function, $p(t|\beta,\mathbf{z})$, or via a hazard rate, $\lambda(t|\beta,\mathbf{z}) = \lambda_0(t)\ e^{\beta\cdot\mathbf{z}}$. Starting from (4.3)-(4.10), we derive some useful identities used throughout this paper

$$\begin{aligned} p(t|\beta,\mathbf{z}) &= \lambda(t|\beta,\mathbf{z})\,S(t|\beta,\mathbf{z}) = \lambda_0(t)e^{\beta\cdot\mathbf{z}-\Lambda_0(t)\,e^{\beta\cdot\mathbf{z}}} \\ \log p(t|\beta,\mathbf{z}) &= \log\lambda_0(t) + \beta\cdot\mathbf{z} - \Lambda_0(t)e^{\beta\cdot\mathbf{z}} \end{aligned} \tag{B.32}$$

Using the short-hand $\xi = \beta\cdot\mathbf{z}$ for the linear predictor

$$\frac{\partial}{\partial\xi}\log p(t|\xi) = 1-\Lambda(t)e^{\xi} \tag{B.33}$$

$$\frac{\partial}{\partial\xi}\,p(t|\xi) = \frac{\partial}{\partial\xi}\left(\lambda(t)e^{\xi-\Lambda(t)\,e^{\xi}}\right) = p(t|\xi)\,[1-\Lambda(t)e^{\xi}] \tag{B.34}$$

$$\frac{\partial}{\partial\xi}p(t|Sy_0,\lambda_0) = S\ p(t|Sy_0,\lambda_0)\,[1-\Lambda(t)e^{Sy_0}] \tag{B.35}$$

Similarly for the functional derivative with respect to the hazard rate:

$$\frac{\delta\log p(t|\xi,\lambda)}{\delta\lambda(s)} = \frac{\delta(t-s)}{\lambda_0(s)} - \Theta(t-s)\,e^{\xi} \tag{B.36}$$

B.8 Symmetry in order parameter equations

We show that the order parameter equations (3.29a)-(3.29f) are invariant under the transformation $r^\star \to -r^\star$. Using $y_0 \to -y_0$ and $r_0 \to -r_0$, $p(t=1|\tilde{S}y_0, r_0) \to p(t=-1|\tilde{S}y_0, r_0)$. From (3.29f), we find $r \to -r$.

Applying these transformations to (3.29c)-(3.29e) does not change the order parameter equations. Hence (3.29c)-(3.29f) are unchanged using these relations allowing us to conclude the order parameters $\tilde{u}, v$ and w will be invariant to sign changes of the true intercept value, r_0. This is confirmed numerically in Section 3.3.

B.9 Choice of covariance matrix

The impact of the spectrum of A is channelled strictly via the following quantities, $s \in \{1,2,3\}$ and $t \in \{1,2\}$

$$\Omega(s,t) = \left\langle \frac{a^s}{(2\eta + \tilde{g}a)^t} \right\rangle \tag{B.37}$$

where the brackets represent averages over the eigenvalue spectrum of $\mathbf{A}$ i.e. $\langle f(a)\rangle = \int \mathrm{d}a\, \rho(a) f(a)$. For $\mathbf{A} = \mathbb{I}_p$, we would find $\Omega(s,t) = \Omega_0(s,t) = (2\eta + \tilde{g})^{-t}$. The four combinations found in the equations order parameter equations are $(s,t) \in \{(1,1),(2,1),(2,2),(3,2)\}$ so $s - t \leq 1$.

Covariance matrix 1. Consider a covariance matrix with the form $A_{\mu\nu} = \delta_{\mu\nu} + (1 - \delta_{\mu\nu})\varepsilon/\sqrt{p}$. It has eigenvalues $1 - \varepsilon/\sqrt{p}$ (with multiplicity $p-1$) and $1 + (p-1)\varepsilon/\sqrt{p}$ (with multiplicity 1), so the first two moments of the eigenvalue spectrum are $\langle a\rangle = 1$ and $\langle a^2\rangle = 1 + \varepsilon^2$. Hence the requirements for self-averaging of the RS theory on the eigenvalue spectrum of $\mathbf{A}$ are fulfilled, viz. $\lim_{p\to\infty}\langle a\rangle < \infty$ and $\lim_{p\to\infty} p^{-1}\langle a^2\rangle = 0$. Calculate the limit $p \to \infty$ of the general form $\Omega(s,t)$

$$\begin{aligned}
\lim_{p\to\infty} \Omega(s,t) &= \lim_{p\to\infty} \left\{ \frac{1}{p} \frac{[1 + (p-1)\varepsilon/\sqrt{p}]^s}{[2\eta + \tilde{g}(1 + (p-1)\varepsilon/\sqrt{p})]^t} + \frac{p-1}{p} \frac{[1 - \varepsilon/\sqrt{p}]^s}{[2\eta + \tilde{g}(1 - \varepsilon/\sqrt{p})]^t} \right\} \\
&= \Omega_0(s,t) + \lim_{p\to\infty} \left\{ \frac{1}{p} \frac{\varepsilon^s\, p^{s/2}\, (1 + \mathscr{O}(p^{-s/2})}{\tilde{g}^t\, p^{t/2}\, (1 + \mathscr{O}(p^{-t/2})} \right\} \\
&= \Omega_0(s,t) + \varepsilon^{s-t} \tilde{g}^{-t} \lim_{p\to\infty} p^{(s-t)/2\, -1}
\end{aligned} \tag{B.38}$$

Since $s-t \leq 1$, the limit for all averages considered is $\lim_{p\to\infty} \Omega(s,t) = \Omega_0(s,t)$

Covariance matrix 2. Our second choice for $\mathbf{A}$ had again $A_{\mu\mu} = 1$ for all μ, but now covariates are correlated in ordered pairs: $A_{\mu,\mu+1} = A_{\mu+1,\mu} = \varepsilon$ for all μ odd, with $A_{\mu\nu} = 0$ for all other $\mu \neq \nu$ (with $0 \leq \varepsilon \leq 1$). This is a block diagonal matrix with $\rho(a) = \frac{1}{2}\delta(a-1-\varepsilon) + \frac{1}{2}\delta(a-1+\varepsilon)$, and the RS order parameters *will* depend on the strength ε of the covariate correlations. The function $\Omega(s,t)$ is now

$$\Omega(s,t) = \frac{1}{2}\frac{(1+\varepsilon)^s}{(2\eta+\tilde{g}(1+\varepsilon))^t} + \frac{1}{2}\frac{(1-\varepsilon)^s}{(2\eta+\tilde{g}(1-\varepsilon))^t} \tag{B.39}$$

Since analytical expressions for all values of $\Omega(s,t)$ are available, these are used in the code to avoid finite size effects. For reference these are all different from the $\mathbf{A} = \mathbb{I}_p$ case when $\varepsilon \neq 0$

$$\begin{aligned}
\Omega(1,1) &= \frac{1}{2}\frac{(1+\varepsilon)}{(2\eta+\tilde{g}(1+\varepsilon))} + \frac{1}{2}\frac{(1-\varepsilon)}{(2\eta+\tilde{g}(1-\varepsilon))} \\
\Omega(2,1) &= \frac{1}{2}\frac{(1+\varepsilon)^2}{(2\eta+\tilde{g}(1+\varepsilon))^1} + \frac{1}{2}\frac{(1-\varepsilon)^2}{(2\eta+\tilde{g}(1-\varepsilon))^1} \\
\Omega(2,2) &= \frac{1}{2}\frac{(1+\varepsilon)^2}{(2\eta+\tilde{g}(1+\varepsilon))^2} + \frac{1}{2}\frac{(1-\varepsilon)^2}{(2\eta+\tilde{g}(1-\varepsilon))^2} \\
\Omega(3,2) &= \frac{1}{2}\frac{(1+\varepsilon)^3}{(2\eta+\tilde{g}(1+\varepsilon))^2} + \frac{1}{2}\frac{(1-\varepsilon)^3}{(2\eta+\tilde{g}(1-\varepsilon))^2}
\end{aligned} \tag{B.40}$$